The Dissertation Committee for Seong-Hyeon Kim Certifies that this is the approved version of the following dissertation:

Investigating Factor Structure of Scores on the Outcome Questionnaire Using Factor Mixture Modeling

Committee:

Alissa R. Sherry, Co-Supervisor

S. Natasha Beretvas, Co-Supervisor

David J. Drum

Christopher J. McCarthy

Jane M. Bost

Investigating Factor Structure of Scores on the Outcome Questionnaire Using Factor Mixture Modeling

by

Seong-Hyeon Kim, B.A.; M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2009

Dedication

This work is dedicated to my dear wife, Seung Jin (Jean) Park, the most precious gift for me from the Eternal Father.

Acknowledgements

I am very fortunate to have a supportive, insightful, and intellectually challenging committee who immensely contributed to this study and my writing. I am particularly grateful to my co-chairs, Alissa Sherry and Tasha Beretvas whose ongoing mentorship and encouragement has been invaluable. I am also indebted to David Drum and Jane Bost, who helped me to appreciate the joy of counseling and ushered me to ask and answer clinical significance and implications of my research. I am also grateful to Chris McCarthy, who has been consistently supportive and accepting throughout my graduate study.

I sincerely thank Chris Brownson at the UT Counseling and Mental Health Center to let me use a dataset collected by the Research Consortium, without which this study could have been impossible. I am grateful to my family for their consistent support, interest, and excitement in my life and work. I am grateful to those who supported me in their prayers: Bennie Crum, Joel Wong, Jerry & June Briney, Kim & Julie Williams, Doris Beseda, Anita Turpin, and the members of Austin Korean Presbyterian Church. I extend my thanks to Kay Hyon Kim, Chang-Dai Kim, and Dong-il Kim at Seoul National University for their inspiration and guidance during my master's years in Korea.

Finally, my special thanks go to my beloved wife, Jean, for her unwavering love, wonderful sense of humor, and consistent belief in me.

Investigating Factor Structure of Scores on the Outcome Questionnaire Using Factor Mixture Modeling

Publication No._____

Seong-Hyeon Kim, Ph.D.

The University of Texas at Austin, 2009

Supervisors: Alissa R. Sherry and S. Natasha Beretvas

The Outcome Questionnaire (OQ-45; Lambert et al., 1996) has been widely employed as a psychotherapy outcome monitoring measure following research findings that support various aspects of its validity and sensitivity to change. Despite its broad usage in both clinical and research settings, some of its psychometric properties are not definite. The three subscales of the OQ-45 are designed to measure three distinct, but related, aspects of psychological functioning. However, neither the one- nor three-factor models have been supported by previous research. Likewise, the results of the current study supported neither of those factor structures.

It was suspected that heterogeneity in data might have led to the lack of the confirmatory factor analysis model fit. Therefore, factor mixture modeling (FMM), a combination of confirmatory factor analysis and latent class analysis, was employed to investigate potential heterogeneity of the data. Among the series of factor mixture models with varying numbers of classes that were fitted, the two-class, unconditional FMM

based on the revised three-factor solution was decided to best describe the data under analysis. Although three covariates of clinical status, sex, and race were selected as known sources of heterogeneity and incorporated into the FMMs (i.e., conditional model), the findings were contradictory to expectations.

The implications of these findings in counseling were discussed in terms of aggregating OQ-45 scores and its score interpretation. Furthermore, this study demonstrates the process involved and dilemmas encountered in choosing the best fitting FMM. There is currently no criterion for assessing individual model fit. Instead, models' fit are compared using various information criteria (IC). And, as was found in the current study, these ICs are frequently contradictory. Thus, the process of identifying the best fitting model cannot rest solely on fit indices but must also depend on interpretation of models and consideration of the ultimate use of the results. In the current study, consideration of transition matrices and the pattern of latent means across classes contributed as much to model selection as fit index interpretation.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The clinical improvement of a psychotherapy client is routinely assessed. Although much of this assessment is still done in a casual manner, standardized outcome assessment tools also are routinely utilized by psychotherapy professionals. The use of standardized outcome measures is expected to grow, given a stronger demand for accountability in psychotherapy (Hatfiled & Ogles, 2004). However, it should be noted that psychotherapy outcomes entail many important issues with regard to ethics, research, philosophy, and theory (Lambert, Ogles, & Masters, 1992). Thus, the kind of outcome measure used in a psychotherapy setting becomes a serious issue. Scores on the assessment tool should possess reasonable, appropriate psychometric properties, such as adequate reliability and validity. The present research aims to investigate the psychometric properties of scores on the Outcome Questionnaire (OQ-45).

The Outcome Questionnaire (OQ-45; Lambert et al., 1996) has been widely adopted as a tool for monitoring treatment efficacy in clinical settings, for making informed decisions about clinically significant change, and for establishing psychotherapy goal criteria (Lambert, Gregersen, & Burlingame, 2004). Hatfield and Ogles (2004) ranked the OQ-45 as one of the most extensively used psychotherapy outcome measures. Furthermore, the OQ-45 is one of the most widely used research measures in the field of psychology. For example, the Research Consortium, a nationwide organization comprising over 40 college counseling centers, has been using it to enhance studies on the mental health status of college students. Last, the OQ-45 has been translated into German (Lambert, Hannöver, Nisslmüler, Richard, & Kordy, 2002)

and a few modified versions of the OQ-45 are currently in use (Brown, Burlingame, Lambert, Jones, & Vaccaro, 2001; Brown, Lambert, Jones, & Minami, 2005; Burlingame, Wells, Lambert, & Cox, 2004; Dunn, Burlingame, Wells, Walbridge, Smith, & Crum, 2005). For example, the Life Status Questionnaire (LSQ), a 30-item version constructed based on the OQ-45, has been utilized by PacifiCare Behavioral Health, Inc., a managed behavioral health care organization serving more than 4.1 million members across nine western states (Brown, Burlingame, Lambert, Jones, & Vaccaro, 2001; PacifiCare Behavioral Health, Inc., 2005).

In spite of its extensive application in both clinical practice and research, some of the psychometric properties of the OQ-45 are not well supported. First, its supposed factor structure has never been empirically supported. The three subscales of the OQ-45 designed to measure three distinct, but related, aspects of psychological functioning imply the existence of three factors. On the other hand, high correlations among the three subscales suggest a single overall factor denoting psychological distress (Lambert, Gregersen, & Burlingame, 2004). However, neither of these factor models was given decent support in the only two extant studies designed to assess the psychometric properties of the OQ-45 (Beretvas & Kearney, 2003; Mueller, Lambert, & Burlingame, 1998). Mueller, Lambert, & Burlingame (1998) attempted a confirmatory factor analysis (CFA) to provide support for the factorial validity of the OQ-45. They fitted one-, two-, and three-factor models on the OQ-45 scores, but none of these models showed a decent fit to the data. Given these inauspicious results, they recommended further confirmatory factor analyses on the scores of the OQ-45, by which a reasonable and consistent factor model will be identified and, also, validated among a number of groups. Following the

recommendation, Beretvas and Kearney (2003) attempted another CFA study, and specified and tested one- and three-factor models with all 45 items, but neither fitted the data. This led them to get rid of 24 items with either low factor loadings or low relevancy to other items measuring the same factor. However, even this shortened version with only 21 items measuring four factors of Depression/Anxiety, Feeling of Well-Being, Impact of Stress, and Alcohol Abuse fitted the data only marginally well.

However, contrary to their expectations, poor model fit could arise not only from a lack of a reasonable and consistent factor model but from the possibility that the data under investigation originated from heterogeneous populations with different parameters such as discrepant factor loadings and error variances (Jedidi, Jagpal, & DeSarbo, 1997). In fact, one fundamental assumption of CFA is that data come from a single, homogeneous population and, thus, one covariance matrix is used to represent the intercorrelations of the variables (Bauer and Curran, 2004). Mueller, Lambert, and Burlingame's (1998) and Beretvas and Kearney's (2003) CFA studies were carried out based on this assumption. For example, it is plausible that the true factor model for the OQ-45 is not a single class model but a multi-class model where the three subscales match three factors as expected, but the values of the loadings, for example, might differ across classes. If this is true, this would explain some of the lack of fit identified in the previous OQ-45 CFA studies. However, some may argue that multi-group CFA (MG-CFA) can be used as an alternative to overcome CFA's homogeneity assumption because MG-CFA allows for an analysis of multiple samples at the same time. Nevertheless, MG-CFA requires that the different groups be clearly specified prior to analysis. Hence, MG-CFA becomes useless when the source of heterogeneity is unobserved.

In addition to the questionable factor validity of the OQ-45, research comparing the mean OQ-45 scores for groups such as sex or race has been inconclusive so far. Despite these inconclusive results (Lambert, Hannöver, Nisslmüller, Richard, & Kordy, 2002; Lambert, Smart, Campbell, Hawkins, Harmon, & Slade, 2006; Nebeker, Lambert, & Huefner, 1995), the OQ-45 developers argue there is no need to provide separate norms for different groups such as males and females (Lambert, Gregersen, & Burlingame, 2004). However, group differences on psychological measures are important to understand because sometimes any significant group difference could originate from measurement bias inherent in the measure, rather than from true differences between the groups of interest (Allen & Yen, 1979).

ANOVA, which has mainly been used to examine group mean differences on the OQ-45, can only be used to compare groups defined by observed variables such as sex and race on total scores for a scale. In other words, ANOVA cannot incorporate any type of latent variables in its framework (Cole, Maxwell, Arvey, & Salas, 1993). ANOVA does not permit a comparison that controls for measurement error. With regard to the undesirable effects of measurement error in ANOVA, Hancock (1997, 2004) argues that measurement error affects both the numerator and the denominator in the *F*-ratio of $MS_{between} / MS_{within}$, distorting the true relationship of between-group variance and within-group variance. This often leads to questionable results. This same criticism applies for other group comparison techniques based on observed grouping variables such as MANOVA, discriminant analysis, and logistic regression.

Therefore, given the inconsistent findings in terms of the factor structure on OQ-45 scores and of observed group mean differences on the OQ-45, it is suggested (Lubke & Muthén, 2005) that psychometric properties of the OQ-45 be investigated using a more sophisticated method that successfully controls the methodological limitations of CFA and ANOVA. Unlike these more traditional methods of CFA and ANOVA, an alternative statistical method should be based on a latent variable system such as factor mixture modeling (FMM).

Factor mixture modeling (FMM), which integrates both continuous and categorical latent variables in its framework, can be utilized as an alternative to CFA and ANOVA. The next section will discuss more about the characteristics and purposes of FMM. Factor mixture modeling is a combination of confirmatory factor analysis and latent class analysis (LCA) where LCA is a form of mixture modeling (discussed more thoroughly in chapter 2). CFA and LCA are both latent variable models in which either continuous or categorical latent variables are assumed to explain the covariation (i.e., correlation or covariance) among a group of observed variables (Bauer & Curran, 2004; Borsboom, Mellenbergh, & Heerden, 2003; Muthén, 2002). FMM mainly serves two purposes by combining the advantages of LCA and CFA, namely: a) classifying groups of people without using observed grouping variables such as sex or race, and (b) investigating factor structure and factor mean differences across classes. CFA attempts to explain covariation among a set of observed variables with a continuous latent variable assumed to commonly affect those observed variables. On the other hand, LCA is a statistical modeling technique with categorical latent variables that represent discrete subpopulations or classes in a population (Muthén & Muthén, 1998-2006). Although

LCA tries to explain all the covariance among the observed variables, it assumes, contrary to CFA, that the shared variation manifests the existence of discrete subpopulations in the population (Bauer & Curran, 2004).

Despite its flexibility for modeling complex distributions, factor mixture modeling has not been widely adopted in psychology research. Only recently in the field of psychology has an extension of factor mixture modeling, namely, growth mixture modeling been seen to be used (e.g., Colder, Campbell, Ruel, Richardson, & Flay, 2002; Colder, Mehta, Balanda, Campbell, Mayhew, Stanton, Pentz, & Flay, 2001; Ellickson, Martino, & Collins, 2004; Greenbaum, Del Boca, Darkes, Wang, & Goldman, 2005; Jackson & Sher, 2005; McCullough, Enders, Brion, & Jain, 2005; Orlando, Tucker, Ellickson, & Klein, 2004; Reinecke, 2006; Stoolmiller, Kim, & Capaldi, 2005; 2001; Tucker, Orlando, & Ellickson, 2003; White, Bates, & Buyske, 2001). Mixture models have mainly been employed in studies in biology, astronomy, and genetics in the natural sciences, and marketing and economics in the social sciences (McLachlan & Peel, 2000). When FMM is used to examine differences in factor structures and factor means in different classes, FMM can be conceptualized as consisting of several steps. First, if an appropriate factor model is defined by theory or in previous research, the relevant confirmatory factor model is estimated for a single class and the potential for additional classes is explored. If a number of classes are identified (using both statistical and substantive evidence), then the factor structures and factor means of the classes can be compared. Last, interpretation of what distinguishes the classes can be investigated through the addition of covariates to the model. This same procedure will be followed using scores on the OQ-45 as well as observed characteristics of participants. The present

study will demonstrate the strengths and flexibility of FMM for use in validating scores on psychological measures.

The current study purports to address the weaknesses of previous OQ-45 validity studies by using FMM. This study has three closely related goals. First, it is designed to identify an appropriate factor structure for OQ-45 scores, which has not been accomplished in previous research (Beretvas & Kearney, 2003; Mueller, Lambert, & Burlingame, 1998). For this goal, a confirmatory factor analysis and, if necessary, an additional exploratory factor analysis will be attempted. Second, the present research aims to assess whether there is any heterogeneity in the factor structures and factor means across unobserved subpopulations through the use of factor mixture modeling. Third, the present study attempts to investigate the relationship between subjects' observed characteristics that may introduce heterogeneity and their latent class membership by incorporating three covariates, specifically: sex, race, and clinical status (i.e., a clinical group and a non-clinical group) into the factor mixture model supported in the previous step.

Chapter 2: Review of the Literature

THE OUTCOME QUESTIONNAIRE (OQ-45)

In response to the demand for a reliable measure of psychotherapy progress monitoring, the OQ-45 (Lambert et al., 1996) was developed as a brief assessment for a wide range of outpatient settings. Since its initial publication in 1994 (Lambert, Lunnen, Umphress, Hansen, & Burlingame, 1994), its utilization among psychologists as a psychological treatment outcome measure has so rapidly grown that it has become one of the most frequently used psychotherapy outcome measures (Hatfield & Ogles, 2004).

The OQ-45 is a self-report instrument designed to assess problems common to a wide variety of adult mental disorders and syndromes and to be employed as a baseline evaluation tool in primary care for referral for psychological therapies (Lambert, Gregersen, & Burlingame, 2004). The OQ-45 instructions direct respondents to answer the items on the basis of how they have felt over the past week. The instrument consists of 45 items, all of which are based on a five-point Likert scale, including values of 0 (*never*), 1 (*rarely*), 2 (*sometimes*), 3 (*frequently*), and 4 (*almost always*). To decrease the possibility of obtaining biased results arising from response sets, the OQ-45 was constructed so that increasing scores correspond to increasing levels of psychopathology on 36 of the OQ-45 items (e.g., "I feel no interest in things"), whereas increasing scores correspond to decreasing levels of psychopathology on nine of the OQ-45 items (e.g., "I get along well with others"). These nine positive items are reverse scored to get the total score. The total score of the OQ-45 ranges from 0 to 180, with higher scores representing

more frequent, more severe psychological distress, interpersonal problems, less adaptive social functioning, and less frequent positive emotional states.

The OQ-45 is purported to measure (a) symptom distress, (b) interpersonal relations, and (c) social role performance (Mueller, Lambert, & Burlingame, 1998). The three subscales of the OQ-45—Symptom Distress (intrapsychic functioning, e.g., "I feel blue"), Interpersonal Relations (e.g., "I feel lonely"), and Social Role Performance (e.g., "I feel stressed at work/school") are aimed at assessing these three different domains of client functioning (Lambert et al., 1996; Whipple et al., 2003). The Symptom Distress subscale consists of 25 items that evaluate psychological symptoms associated with the most prevalent types of mental disorders among adults (e.g., anxiety disorders, mood disorders, and substance-related disorders). The Interpersonal Relations subscale consists of 11 items that attempt to assess functioning in interpersonal relationships. The last subscale, Social Role Performance, consists of nine items that assess an individual's current level of social role performance (Vermeersch et al, 2004). The three subscales and all the items were "rationally" selected based on relevant literature review. For example, the selection criteria for items in the Symptom Distress subscale were a) conformation to DSM-III-R diagnosis criteria, b) continued reference in research as symptoms of the chosen psychopathology, and c) item analysis (e.g., inter-item correlations) results. After preliminary development of items for each subscale, an item analysis was conducted to investigate reliability of scores. Dubious items, for example item scores with low reliability, were either dropped or changed (Lambert et al., 1996).

Scores on the OQ-45 have exhibited reasonable reliability (Lambert et al., 1996). Three-week test–retest reliability was estimated to be .84 using total OQ-45 scores of 157

undergraduate students. Also, Cronbach's α was estimated to be .93 using total OQ-45 scores of 157 undergraduate students and 298 clinical patients. In addition, the OQ-45 has been demonstrated to be sensitive to change in individuals undergoing psychotherapy over short time periods while remaining stable in untreated populations (Vermeersch et al., 2000).

Validation

Validity, in general, is related to the accuracy and appropriateness of inferences that are made from examinees' responses on a test (Kane, 2006). This viewpoint is clearly evident in the following definition of validity presented in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999, p. 9): "the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests… The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations."

Validity is an umbrella term that incorporates various types of validity such as construct validity and content validity. The validity of scores on the OQ-45 have been extensively researched. For example, scores on the OQ-45 are reported to have high concurrent validity with scores on a variety of self-report scales, such as the Symptom Checklist-90-Revised (SCL-90-R, Derogatis & Cleary, 1977; Derogatis, Rickels, & Rock, 1976), the Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), the Zung Depression Scale (Zung, 1965), Zung Anxiety Scale (Zung, 1971), the State–Trait Anxiety Inventory (Spielberger, 1983; Spielberger, Gorsuch, & Lushene, 1970), the Taylor Manifest Anxiety Scale (Taylor, 1953), the Inventory of

Interpersonal Problems (IIP, Horowitz, Rosenberg, Baer, Ureño, & Villaseñor, 1988), and the Social Adjustment Scale (Weissman & Bothwell, 1976) (Lambert et al., 1996; Umphress, Lambert, Smart, Barlow, & Clouse, 1997).

Nevertheless, among various kinds of validity, construct validity is presently considered as the entirety of validity theory overarching all other types of validity such as content and criterion (e.g., predictive and concurrent validity) validity (Kane, 2001; Zumbo & Rupp, 2004). According to Messick (1998), these other types of validity cannot operate independently but only provides complementary information to construct validity. Construct validity refers to the extent to which inferences made from individuals' responses on a test appropriately captures the theoretical construct the test is intended to measure (Zumbo & Rupp, 2004). Factorial validity, a kind of construct validity, was coined by Guilford (1946, p. 428) prior to the introduction of the concept of construct validity in the 1950s: "The factorial validity of a test is given by its loadings in meaningful, common, reference factors. This is the kind of validity that is really meant when the question is asked 'Does this test measure what it is supposed to measure?'... The answer then should be in terms of factors and their loadings." With regard to factorial validity, a type of construct validity established by factor analysis, scores on the OQ-45 have been shown to be deficient as evidenced by lack of support for its supposed one- and three-factor solutions in two confirmatory factor analysis studies (Beretvas & Kearney, 2003; Mueller, Lambert, & Burlingame, 1998), which behooves further investigation of what scores on the OQ-45 measure.

Actually, it has been known that the uniqueness of the constructs represented by the three subscales of the OQ-45 is poorly supported (Lambert & Hawkins, 2004). For

example, intercorrelations among scores on three clinical samples' three OQ-45 subscales were examined to evaluate the subscales' independence from one another (Umphress, Lambert, Smart, Barlow, & Clouse, 1997). Seventeen of the 18 correlations were moderate to strong (ranging from .42 to .98) and statistically significant indicating that the subscales share considerable variance. The study also demonstrated that most of the variance in the OQ-45 total score was best explained by the Symptom Distress subscale with the highest subscale-total score correlations of .96, .96, and .98 among the nine subscale-total score correlations from the three clinical samples. This suggests that the OQ-45 is best considered as a measure of subjective psychological distress. The predominantly strong intercorrelations between each of the OQ-45 subscales (average correlation of .59 across three clinical samples) indicate that the subscales are not measuring discrete domains of psychological functioning (Lambert, Gregersen, & Burlingame, 2004).

Mueller, Lambert, and Burlingame (1998) conducted a confirmatory factor analysis (CFA) of the OQ-45 as an attempt to validate its purported multi-factor structure. However, the study failed to support the three dimensions of the OQ-45. In the study all of the models (three-, two-, and one-factor) evaluated did not fit the samples' data at all. Despite this, the authors of this CFA study recommended using only the total score in practice due to the confirmed limited interpretive value of each subscale score. They also recommended additional CFA studies in which empirical and rational manipulation of item pools should be utilized to find a scale that might better fit the proposed three-dimensional structure for the OQ-45 (Lambert, Gregersen, & Burlingame, 2004; Mueller,

Lambert, & Burlingame 1998). These recommendations imply the factorial validity of the OQ-45 is not supported by the data.

Following the recommendation, Beretvas and Kearney (2003) attempted another CFA study to examine the factorial validity of OQ-45 scores. They specified and tested one- and three-factor models with all 45 items, but neither fit the data. An intercorrelated seven-factor model, which was specified after a content analysis on items, was tested, but it also was not found to fit the data. This led them to get rid of 24 items with either low factor loadings or low relevancy to other items measuring the same factor. However, this shortened version with only 21 items measuring four factors of Depression/Anxiety, Feeling of Well-Being, Impact of Stress, and Alcohol Abuse fit the data only marginally well. The four-factor, 21-item model was cross-validated across Hispanic, Asian-American, and African-American samples. Also, a multi-group CFA in which factor loadings and intercorrelations were constrained to be equal across White and Hispanic samples was conducted and it showed an adequate fit to the data supporting the validity of the shortened version in cross-cultural settings. This result also supported the invariance of the factor model across the two ethnic groups. Finally, a MANOVA was done to examine group differences on the four factors between Hispanic and White samples. The mean of Hispanic subjects on the Impact of Stress scale was found to be significantly higher than that of White subjects. The authors recommended further research on the factor structure of scores on the OQ-45 and on the invariance of the factor structure across different groups. However, so far group comparisons (e.g., different sex, age, and ethnic groups) of OQ-45 scores have only involved comparison of raw scores.

No studies have employed a comparison of factor scores for which measurement error is controlled.

It has also been argued that probable consequences from the use of a test should also be taken into account when validating scores on a test (AERA, APA, & NCME, 1999; Kane, 2001). Nevertheless, consequential validity has not been given due attention in the discussion of validation (Kane, 2001), and some believe the concept of consequential validity is not necessary because considering an extra aspect of social consequences of test use inevitably induces confusion (Popham, 1997). However, it would be beneficial to investigate the OQ-45 in terms of its consequential validity.

Scores on the OQ-45 are used to describe (i.e., diagnose) and categorize people into, for example, clinical and non-clinical groups and are also used to prescribe corresponding adequate treatments based on clinical decisions. In addition, mixed results have been reported in terms of racial differences in OQ-45 scores. Caucasians, Pacific Islanders, and Asians were found to have significant differences on the OQ-45 total scores (Gregersen, Nebeker, Seely, & Lambert, 2004). Given the weak factorial validity of OQ-45 scores, a question still remains: whether the differences among these racial groups are due to the differences on the construct or to the differences on something unrelated to the construct. This question is important because any differences between racial groups imply different decision-making and interventions strategies. This demonstrates how the consequential validity of the OQ-45 can be attenuated by its weak factorial validity.

Nevertheless, there is an indication that the weak OQ-45 factorial validity undermines not only its consequential validity but also its convergent and discriminant

validity. A thorough re-evaluation of concurrent correlation coefficients reported in the two validity studies (Lambert et al., 1996; Umphress, Lambert, Smart, Barlow, & Clouse, 1997) shows that little support has been found for its convergent and discriminant validity. Before these two validity studies were conducted (Lambert et al., 1996; Umphress, Lambert, Smart, Barlow, & Clouse, 1997), the authors of the studies had expected the OQ-45 subscale scores would be highly correlated with their associated criterion measures, manifesting high convergent validity. It was hypothesized the Symptom Distress (SD) subscale would show a strong correlation with the SCL-90-R's General Severity Index (GSI). Also, it was expected the Interpersonal Relations (IR) subscale would be highly associated with the Inventory of Interpersonal Problems (IIP, Horowitz, Rosenberg, Baer, Ureño, & Villaseñor, 1988), a 127-item self-report measure that inventories interpersonal distresses experienced in significant relationships. Last, the Social Role (SR) subscale was expected to be highly correlated with the Social Adjustment Scale (SAS, Weissman & Bothwell, 1976), a 56-item self-report questionnaire that measures social role functioning in several domains. In fact, at least moderate levels of correlations were found between the OQ-45 subscales and their corresponding criterion measures testifying to its convergent validities (see Table 1). Interestingly, Table 1 can be used as a form of multi-trait multi-method matrix (MTMM), which was developed by Campbell and Fiske (1959) as a way of investigating convergent and discriminant validity simultaneously. However, the correlation coefficients (r = .49, .64, and .57) between the IR subscale scores and the IIP scores are lower than those (r = .60, .70, and .86) between the SD subscale scores and the IIP score. Note that the former correlations are supposed to be higher than the latter. If the IR subscale is supposed to

measure what the IIP measures (i.e., interpersonal relationship) measures as the authors expected, the correlation between the IR and the IIP should be higher than the other correlation between the IR and the other two criteria (i.e., SCL-90-R's GSI and SAS) that measure something other than interpersonal relationship. This unexpected pattern of correlations was shown again with the OQ-45 SR subscale across the three different samples (see Table 1). For example, the correlation between the SR and the SAS in the inpatient sample was .53, but the correlation between the SD and the SAS was .79, contrary to expectation. If the SAS measures a construct similar to what the OQ-45 SR measures (i.e., social role functioning) but, supposedly, not something close to what the OQ-45 SD subscale measures (i.e., symptom distress), the correlation between the SR and the SAS should be higher than the correlation between the SD and the SAS. Nevertheless, in the Table 1, these hypothesized associations are manifested in a reversed way, indicating that perhaps the OQ-45 subscales do not discretely measure what they are supposed to measure and they are highly confounded (i.e., weak factorial validity). This reversed association also shows how the weak factorial validity of the OQ-45 reduces its convergent and discriminant validity.

Table 1
Validity Estimates for the Outcome Questionnaire

Criterion	OQ total score	Symptom distress	Interpersonal relations	Social role
GSI[2]	0.78 (*N* = 115)[1]	0.61 (*N* = 115)[1]		
	0.72 (*N* = 238)	0.70 (*N* = 238)	0.50 (*N* = 238)	0.52 (*N* = 238)
BDI[2]	0.80 (*N* = 115)[1]	0.63 (*N* = 115)[1]		
	0.62 (*N* = 238)	0.59 (*N* = 238)	0.44 (*N* = 238)	0.47 (*N* = 238)
ZSDS[2]	0.88 (*N* = 71)	0.89 (*N* = 71)	0.67 (*N* = 71)	
ZSAS[2]	0.80 (*N* = 71)	0.81 (*N* = 71)	0.53 (*N* = 71)	0.71 (*N* = 71)
TMAS[2]	0.86 (*N* = 71)	0.88 (*N* = 71)	0.63 (*N* = 71)	0.64 (*N* = 71)
STAI(y-l)[2]	0.64 (*N* = 115)[1]	0.50 (*N* = 115)[1]		
STAI(y-2)[2]	0.80 (*N* = 115)[1]	0.65 (*N* = 115)[1]		
IIP[2]	0.60 (*N* = 71)	0.60 (*N* = 71)	0.54 (*N* = 71)	0.47 (*N* = 71)
	0.63 (*N* = 238)	0.58 (*N* = 238)	0.50 (*N* = 238)	0.60 (*N* = 238)
SAS[2]	0.62 (*N* = 71)	0.56 (*N* = 71)	0.65 (*N* = 71)	0.44 (*N* = 71)
	0.60 (*N* = 238)	0.52 (*N* = 238)	0.47 (*N* = 238)	0.41 (*N* = 238)

Note. [1]These values were obtained with a preliminary 43 item version of the current 45 item test. [2]GSI = General Severity Index; BDI = Beck Depression Inventory; ZSDS = Zung Self Rating Depression Scale; ZSAS = Zung Self Rating Anxiety Scale; TMAS = Taylor Manifest Anxiety Scale; STAI = State Trait Anxiety Inventory (y-1 = State Anxiety; y-2 = Trait Anxiety); IIP = Inventory of Interpersonal Problems; SAS = Social Adjustment Scale. From "The reliability and validity of the Outcome Questionnaire," by M. J. Lambert, G. M. Burlingame, V. Umphress, N. B. Hansen, D. A. Vermeersch, G. C. Clouse, and S. C. Yanchar, 1996, *Clinical Psychology and Psychotherapy, 3*, p. 255.

Also, one may be interested in how the weak factorial validity of OQ45 scores can affect clinical practices and research findings. It would give us insights into this issue if we consider how other psychological measures with weak factorial validity are regarded by researchers. For example, in line with the OQ-45, the SCL-90-R (Derogatis & Cleary, 1977; Derogatis, Rickels, & Rock, 1976) has been known to suffer from weak factorial validity in spite of its wide acceptance. Even though the SCL-90-R is supposed to comprise nine dimensions (i.e., subscales), no confirmatory factor analysis has been

done to verify its factor structure. In addition, several exploratory factor analysis studies have produced factor structures with as low as three up to twelve factors (Arrindell, Barelds, Janssen, Buwalda, & van der Ende, 2006). With regard to the weak factorial validity of the SCL-90-R, Barkham et. al.'s (1998) comment is well worth mentioning. "The implication is that using the design[ed] subscales may not be valid and may give the impression of measuring a number of distinct state dimensions when in fact these are highly confounded. If so, then researchers and practitioners are left with using a new set of scales… Alternatively, users might utilise just the global score parameters based on the full set of 90 items to derive an overall estimate of distress. However, it seems unlikely that 90 items are required to arrive at such a summary estimate." By the same token, only the total score of the OQ-45 is recommended for use because the three-dimension model was not supported (Lambert, Gregersen, & Burlingame, 2004; Mueller, Lambert, & Burlingame, 1998). However, it should be noted that no evidence has been found supporting the unidimensional model (i.e., representing general psychological functioning) (Beretvas & Kearney, 2003; Mueller, Lambert, & Burlingame, 1998) and accordingly the use of the total score is not also warranted. In other words, for example, even though a client shows improvements on the OQ-45 total score, it does not necessarily mean the client's general functioning also got better because we don't know what the OQ-45 exactly measures (Beretvas & Kearney, 2003; Mueller, Lambert, & Burlingame, 1998). This also implies poorly supported score interpretations may result in unreasonable clinical decisions and interventions. Remember what the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999, p. 9) says with regard to this, "The process of validation involves accumulating evidence to provide a

sound scientific basis for the proposed score interpretations." It is believed that we do not have well-founded scientific evidence for an appropriate score interpretation for the OQ-45 and this study aims to provide a "sound scientific basis" for OQ-45 score interpretation.

However, the rationale for this study does not necessarily mean to negate the positive assets of the OQ-45 that have been supported throughout empirical studies. For example, the OQ-45 is unique in that it was developed as a psychotherapy outcome measure for a use in the era of managed health care. Also, the OQ-45 was designed to a large extent for convenient scoring and minimal cost (Umphress, Lambert, Smart, Barlow, & Clouse, 1997). In addition, it has been shown that the OQ-45 is sensitive to psychological changes over a brief period of time and capable of measuring an extensive array of psychological functioning (Umphress, Lambert, Smart, Barlow, & Clouse, 1997; Vermeersch et al., 2000). Its 45 items were mainly developed from the literature that assesses three content domains of symptom distress, interpersonal relations, and social role functioning. These three areas were chosen to be significant for evaluating clients' psychological status and measuring psychotherapy outcome (Umphress, Lambert, Smart, Barlow, & Clouse, 1997). Based on this, although one may say the OQ-45 possesses enough content validity, its factorial validity is still in question. Note Guilford (1948) said that investigating the factorial validity of a test is asking, "Does this test measure what it is supposed to measure?" Thus, we are not sure the OQ45 really measures what it is supposed to measure.

Clinical Usage and Decision-making

Cutoff scores of 63 and 64 were suggested to differentiate a non-clinical or functional population and a clinical or dysfunctional population. These cutoff scores were computed based on a formula developed by Jacobson and Truax (1991) using the OQ-45 normative data from a non-clinical group and a clinical group (Hannan, Lambert, Harmon, Nielsen, Smart, Shimokawa, & Sutton, 2005). If a person's OQ-45 total score falls 63 or below, he or she is more likely to be a member of the functional population. Higher scores place one in the clinical range.

In addition, a reliability change index (RC or RCI) was devised also based on the same formula by Jacobson and Truax (1991). The formula can be expressed as: $RC = (x_2 - x_1) / S_{diff}$, where x_1 indicates a subject's pretest score, x_2 indicates the same subject's posttest score, and S_{diff} denotes the standard error of the difference between the two scores. The RCI indicates how much change should occur for a person to be regarded as having undergone a clinically significant change and was estimated using the normative data to be 14 points. Thus, if the OQ-45 total score of a person in the clinical group changes by more than 14 points and drops below 64, he or she is regarded as meeting the criteria for recovery (Hannan et al., 2005; Lambert, Gregersen, & Burlingame, 2004).

The RCI also was used to categorize different levels of psychotherapy outcome. For example, four different categories were delineated by Hansen, Lambert, and Forman (2002): (a) *deteriorated*-a person's OQ-45 total score has reliably (i.e., decreased more than 14 points) changed into a negative direction over the course of treatment, (b) *no*

change-the score has not varied more than 14 points in any direction, (c) *improved*-the score has changed more than 14 points in the positive direction during treatment, and (d) *recovered*-the score has changed more than 14 points and, at the same time, the person moved from the dysfunctional state to the functional state (i.e., the OQ-45 total score has dropped below 64 from equal or higher than 64).

The OQ-45 has also been utilized as a clinical decision making tool to improve psychotherapy outcomes by monitoring patients' progress and treatment response. For example, Lambert, Hansen, and Finch (2001) devised an expected recovery rate for groups of patients with different initial levels of symptom distress measured with the OQ-45 and identified patients who showed less than satisfactory treatment results. These patients were called signal cases. When feedback on patients' progress was given to the therapists of signal cases, the patients' treatment outcome results were found to be more enhanced than the patients of the corresponding therapists who did not receive feedback. Also, Lambert, Whipple, Bishop, et al. (2002) developed a more rational feedback algorithm, where decision rules were created based on patients' initial distress level measured with the OQ-45, the number of sessions each patient completed, and the amount of change at the session of interest compared with the initial OQ-45 score. The decision rules generate four distinct types of feedback with a progress report and each type of feedback is assigned a different color. For example, whereas a green feedback indicates the patient is making decent progress and no treatment change is recommended, red feedback indicates the patient is not progressing as expected and the patient might be in danger of premature dropout. Identification of a red feedback typically results in a treatment plan review, referral, or treatment intensification.

A further meta-analysis as well as an additional study provided support for the decision-making system where feedback based on patients' progress monitoring increases positive treatment outcomes and prevents treatment failure (Lambert, Whipple, Hawkins, Vermeersch, Nielsen, & Smart, 2003; Whipple, Lambert, Vermeersch, Smart, Nielsen, & Hawkins, 2003). Use of OQ-45 scores as a feedback mechanism is one example of how the OQ-45 is used in clinical settings. Therefore, it is important that the measurement structure of the OQ-45 is carefully assessed. Otherwise, decisions based on the measure become more or less arbitrary rather than clinically informative.

Utilization in Research

In addition to its wide acceptance in clinical settings, the OQ-45 and its shortened, 30-item version have been used in research studies to examine therapist effects (Okiishi, Lambert, Nielsen, & Ogles, 2003; Wampold & Brown, 2005), and the dose-effect model in naturalistic settings (Hansen, Lambert, & Forman, 2002).

First, the OQ-45 and its related version were utilized to clarify the proportion of psychotherapy outcome variability due to therapists. Okiishi, Lambert, Nielsen, and Ogles (2003) analyzed OQ-45 scores of 1,779 college students seen by 56 therapists at a large university counseling center in order to investigate if some therapists produced better outcomes among their clients. They found that there was a significant variation among therapists in their clients' psychotherapy outcome measured by the OQ-45. For example, the clients of the most effective therapist improved ten times faster than the average progress rate of the sample. In a similar vein, Wampold and Brown (2005) examined the extent of psychotherapy outcome variation attributable to individual

therapists. In order to explore their research question, they used a psychotherapy outcome dataset collected by a managed behavioral health care organization. The outcome measure was a shortened, 30-item version of the OQ-45, called the Life Status Questionnaire (LSQ). The final sample for analysis included 6,146 patients treated by 581 therapists. The results revealed that about 5% of the outcome variation is attributed to individual therapists after controlling for initial symptom levels of patients.

Second, the OQ-45 was utilized as a tool to verify the dose-response effect model, which suggests there is a positive association between the number of psychotherapy sessions and the extent of improvement within patients. Hansen, Lambert, and Forman (2002) used a naturalistic dataset of 6,072 patients who completed the OQ-45 prior to each session. The results supported the dose-response effect model in general that more sessions produce better outcome. However, it was revealed that most of the patients in the sample did not receive enough doses for a moderate level of improvement.

Other Related Versions of the OQ-45

A few related but modified versions of the OQ-45 are also available. The Youth Outcome Questionnaire (Y-OQ) was designed in 1999 as a parent-report measure to monitor adolescents' psychotherapy outcome progress. The Y-OQ consists of 64 items comprising six subscales of Interpersonal Distress, Somatic, Interpersonal Relations, Critical Items, Social Problems, and Behavioral Dysfunction, and possible scores range 0 to 256. Each item is rated on a five-point Likert scale. The internal consistency estimates from the normative sample ($N = 467$) for the Y-OQ subscales ranged from .74 to .93 and the estimate for the total score was .96. The Y-OQ also showed high discriminant validity

evidenced from the extensive mean score differences among inpatient, outpatient, and community samples. In addition, the subscales on the Y-OQ demonstrated strong correlations with those of frequently utilized psychological measures for children, such as the Child Behavior Checklist (Burlingame, Wells, Lambert, & Cox, 2004; Wells, Burlingame, Lambert, Hoag, & Hope, 1996).

In addition, Lambert, Burlingame, and their colleagues also developed shortened versions of the OQ-45 and the Y-OQ, namely the Life Status Questionnaire (LSQ) and the Youth Life Status Questionnaire (YLSQ) or Y-OQ-30.1 (Brown, Burlingame, Lambert, Jones, & Vaccaro, 2001; Wampold & Brown, 2005). The LSQ and YLSQ each consist of 30 items and thus take less time to administer. The PacifiCare Behavioral Health (PBH) company uses these two self-report questionnaires for its outcomes management program (Brown, Burlingame, Lambert, Jones, & Vaccaro, 2001). PBH is a managed behavioral health care organization serving more than 4.1 million members across nine western states (Brown, Burlingame, Lambert, Jones, & Vaccaro, 2001; PacifiCare Behavioral Health, Inc., 2005). PBH encourages its network of psychotherapy providers to assess outcome regularly using the LSQ for their patients as a part of clinical outcome management (Brown, Lambert, Jones, & Minami, 2005; Wampold & Brown, 2005). Once the service is authorized for a patient, PBH mails a packet of LSQ questionnaires to the therapist, who then administers it at the first, third, and fifth sessions, and at every fifth session thereafter. The patient fills out a form prior to each required session and the therapist faxes it to PBH to have it scored. PBH in return provides test scores and responses on critical items for critical cases, and also informs on patients with satisfactory outcomes (Brown, Lambert, Jones, & Minami, 2005). Some

incentives for submitting the completed LSQ are granted to the therapist, such as automatic authorization of some cases without regard to the score on the LSQ (Brown, Lambert, Jones, & Minami, 2005). The usage of the LSQ has increased substantially in recent years. For example, 70% of clients who were treated by PHB providers completed the LSQ at least once (Wampold & Brown, 2005).

The Y-OQ-30.1 consists of 30 Likert-type items (0 = almost never or never to 4 = almost always or always) from the Y-OQ that were shown to be most sensitive to clinical change. The total score ranges from 0 to 120. The Y-OQ-30.1 can be completed by an adolescent between the ages of 12 and 18 and this version is called the SR Y-OQ-30.1. Also, it can be rated by an observer for youth between the ages of four and 18 and this version of the Y-OQ-30.1 is called the PR Y-OQ-30.1. The Y-OQ-30.1 has six subscales: Somatic, Social Isolation, Aggression, Conduct Problems, Hyperactivity/Distractibility, and Depression/Anxiety. One psychometric validation study for the Y-OQ-30.1 was done based on a sample of community youths (N = 296) and 9,563 Y-OQ-30.1 clinical ratings combined. The internal consistency reliability coefficients for the total score were estimated for both the community and clinical samples to be .92. The same coefficients for the subscales ranged from .55 to .85. Three-week (on average) test-retest reliabilities for the total score ranged from .80 to .91. The overall scores of the community sample were significantly lower than those of outpatient youth, thus providing support for discriminant validity. Nevertheless, no exploratory or confirmatory factor analysis was done to verify its factor structure (Dunn, Burlingame, Walbridge, Smith, & Crum, 2005)

Furthermore, the OQ-45 was translated into German and the German adaptation was standardized on a non-clinical sample of 232 subjects (129 females and 103 males).

Two-week test-retest reliabilities were estimated to be .87 for Symptom Distress, .81 for Interpersonal Relations, .71 for Social Role Performance, and .88 for the total. The internal consistency coefficients were estimated to be .90 for Symptom Distress, .81 for Interpersonal Relations, .59 for Social Role Performance, and .93 for the total. Six intercorrelations among the three subscale scores and the total scores ranged from .59 to .95. The validity of the German adaptation was examined by comparing with the Symptom Checklist 90, the Inventory of Interpersonal Problems, the Social Adjustment Scale, and a Questionnaire on Life-Satisfaction and was found to be adequate. The intercorrelations between the German adaptation total and subscale scores and the other four scales ranged from .45 to .76. However, no exploratory or confirmatory factor analysis was conducted to evaluate its construct validity (Lambert, Hannöver, Nisslmüler, Richard, & Kordy, 2002).

Comparison of the OQ-45 Scores across Groups

Although, at times, differences on OQ-45 total scores between males and females have been found, in general no consistent differences have been found between males and females (Lambert, Gregersen, & Burlingame, 2004). For this reason, the OQ-45 test manual does not provide separate norms for men and women (Lambert, Gregersen, & Burlingame, 2004). For example, statistically significant differences were reported between males and females on the mean scores of the Symptom Distress subscale and the OQ-45 total (Lambert, Hannöver, Nisslmüler, Richard, & Kordy, 2002). However, to the author's knowledge, there have been no published studies dedicated to gender differences

on the OQ-45. In fact, Lambert, Gregersen, and Burlingame (2004) noted the lack of a substantial difference between the two sexes.

Studies comparing different ethnic groups on the OQ-45 total and subscale scores have had even more mixed results. One study based on a large dataset of patient groups did not suggest any meaningful differences on the total score and subscale scores of the OQ-45 across major ethnic groups (Lambert, Gregersen, & Burlingame, 2004), but another study by Gregersen, Nebeker, Seely, and Lambert (2004) reported that there are significant differences among Caucasians, Pacific Islanders, and Asians on the OQ-45 total scores. More specifically, in this study, Asians scored significantly higher than Pacific Islanders, who in turn scored significantly higher than Caucasians. Among different sub-ethnic/cultural groups in the Asian sample, those who identified themselves as Chinese and Koreans scored highest. On the other hand, Nebeker, Lambert, and Huefner (1995) used ANOVA for group comparisons on the OQ-45 total and subscale mean scores and found no significant differences in the mean scores among different ethnic groups. However, in this study, it was noted that African Americans tended to score significantly higher than Caucasians on certain items of the OQ-45; for example, six items from the Symptom Distress subscale, four items from the Interpersonal Relations subscale, and one item from the Social Role subscale.

Given these equivocal results in terms of the OQ-45 total and subscale score differences across the sexes and racial groups, additional research examining group differences on the OQ-45 scores is warranted. Previous research on group mean differences on the OQ-45 scores, mainly based on ANOVA procedures, suffers from at least two methodological weaknesses: (a) comparing groups on observed scores and thus

lacking any of control of measurement error and (b) solely investigating group mean differences based on directly observed grouping variables. More sophisticated modeling techniques could be used to remediate these weaknesses.

In terms of failure to control for measurement error, it has been discussed that ANOVA, the primary statistical model used for comparing observed mean scores, is affected by measurement error (Hancock, 2004; Hancock, Lawrence, & Nevitt, 2000). As a result, the statistical power to detect any group difference at the construct level is weakened (Hancock, 2004; Muthén, 1991). Several authors have discussed the detrimental effect resulting from comparing groups based on observed scores without controlling for measurement error (e.g., see Hancock, 2004; Lubke, Dolan, Kelderman, & Mellenbergh, 2003), providing one of the justifications for latent variable approaches as an alternative (Muthén, 2002). In common factor models, observed scores are decomposed into the portion attributable to the factor and the portion that constitutes measurement error (Meredith, 1993; Meredith & Horn, 2001). Unlike ANOVA, common factor models can be used to compare group means at the construct level free from measurement error.

In addition to the problem related to measurement error, ANOVA also suffers from another limitation. In ANOVA, only clearly defined, observable grouping variables can be explored as the source of group differences. If group differences lie in other than these observed grouping variables, ANOVA cannot detect those differences. True mean differences may lie in categories not readily observed, but in latent categories that could be inferred through response patterns to items on a questionnaire. For example, we may define *psychologically vulnerable* and *psychologically invulnerable* classes and want to

compare their group mean differences on their IQ scores. Here, the main obstacle is how to categorize a group of subjects into *psychologically vulnerable* and *psychologically invulnerable* classes. Obviously, the degree of psychological vulnerability may not be immediately perceived.

It is suspected that these limitations of ANOVA have contributed to the mixed results of group mean comparison on OQ-45 scores. Therefore, it is suggested that true group mean differences on the OQ-45 be investigated with a more sophisticated modeling technique by which these limitations can be overcome. Factor mixture modeling (FMM) is proposed as a substitute to traditional group mean comparison methods such as ANOVA.

Other Outcome Measures

A few studies have been done to investigate psychotherapy outcome measures that have been utilized in research. Froyd, Lambert, and Froyd (1996) identified 334 studies published from 1983 to 1988. Out of 1,430 outcome measures that were used in those studies 851 measures were utilized only once, and of these 851 measures 278 measures were unstandardized and did not report any reliability and validity information. According to them, the 10 most frequently used measures were: the Beck Depression Inventory (in 42 studies), the State-Trait Anxiety Inventory (in 32 studies), weight (in 28 studies), the Hamilton Rating Scale for Depression (in 18 studies), Symptom Checklist-90 and Symptom Checklist-90-R (in 14 studies), Locke Wallace Marital Adjustment Scale (in 13 studies), blood pressure (in 12 studies), heart rate (in 11 studies), and the Minnesota Multiphasic Personality Inventory (in 10 studies). Froyd, Lambert, and Froyd

(1996) stated in the conclusion that effective exchange and presentation of outcome research among clinicians and researchers is hindered due to a lack of systematic outcome assessment.

Another research study analyzed outcome measures adopted in humanistic (i.e., contrary to cognitive-behavioral) psychotherapy research (Levitt, Stanley, Frankel, & Raina, 2005). They analyzed the results of a meta-analysis that identified 116 outcome measures, of which 85 measures were used only once and nine measures were used more than three times. The nine measures were the BDI (Beck et al., 1972; in 12 studies), the Target Complaints (Battle, Imbers, Hoen-Soric, Stone, Nash, & Frank, 1968; in nine studies), the SCL-90-R (Derogatis, Rickels, & Roch, 1976; in eight studies), Dyadic Adjustment Scale (Spanier, 1976; in eight studies), Hamilton Rating Scale for Depression (Hamilton, 1960; in seven studies), the Stait-Trait Anxiety Inventory (Spielberger, Gorsuch, & Lushene, 1970; in six studies), the Couples Therapy Alliance Scale (Pinsoff & Catherall, 1986; in five studies), Hamilton Rating Scale for Anxiety (Hamilton, 1950; in five studies), and Personal Orientation Inventory (Shostrom & Knapp, 1966; in four studies).

There is an outcome measure comparable to the OQ-45 that assesses the global functioning of a client in psychotherapy treatment. The Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM, Evans et al., 2002) was developed in the United Kingdom to assess "core domains of problems" in clients and to routinely evaluate efficacy and effectiveness of psychotherapy. The CORE-OM, a self-report questionnaire, comprises 34, 5-point Likert-type items, rated on a scale of 0 'not at all' to 4 'most or all the time' with a time scale of 'over the last week' (Connell, Barkham, &

Mellor-Clark, 2007). It has four subscales subjective well-being ("I have felt optimistic about my future"), problems/symptoms ("I have felt panic or terror"), life functioning ("I have felt humiliated or shamed by other people"), and risk ("I have thought of hurting myself"). The CORE-OM is recommended for use before and after each session. The reliability and validity of scores on the CORE-OM were examined (Evans et al., 2002), in which the coefficient α for the total and subscale scores ranged from .75 to .94 among clinical and non-clinical samples. Its one-week test-retest reliabilities for the total and subscale scores were estimated to be from .64 to .91 in a non-clinical sample. The concurrent validity was investigated through its correlation with existing psychological measures such as the Beck Depression Inventory (Beck et al., 1961, 1996), the Beck Anxiety Inventory (Beck et al., 1988), and the Brief Symptom Inventory (Derogatis & Melisaratos, 1983), and the total score correlation coefficients ranged from .55 to .88. The correlations among the CORE-OM total and subscale scores were high and ranged from .33 to .93, indicating dominance of a large single factor, which was proved in the following principal component analysis (PCA). However, three components were chosen in the PCA, representing negatively-worded items, risk items, and positively-worded items. The CORE-OM was also proven to be sensitive to change in psychotherapy clients.

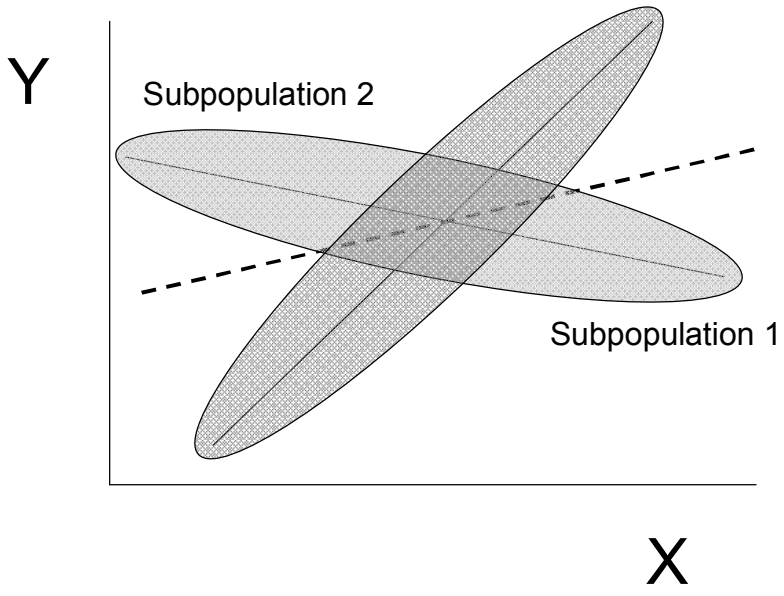
Various aspects of the outcome questionnaire (OQ-45), including its psychometric properties and its related versions of OQ-45 as well as other similar outcome measures, have been explored so far. Remember that the focus of this research is investigating the factorial validity, one of the most important aspects of construct validity, of OQ-45 with a

more advanced statistical method, factor mixture modeling (FMM), which will be extensively discussed in the following section.

FACTOR MIXTURE MODELING

Mixture modeling is designed to deal with a heterogeneous data set, made up of two or more subpopulations that have their own distinctive distributional parameters (McLachlan & Peel, 2000). For example, a data set may contain two distinctive samples that come from two different normal distributions with their own unique means and standard deviations. This data set is called a normal mixture and a normal mixture model should be utilized in order to analyze this type of mixture data. A data set may include two or more subpopulations, for each of which a different factor model can be fitted. This data set can be analyzed with factor mixture modeling. Figure 1 shows a regression mixture model or a mixture of regression analysis (Muthén & Muthén, 1998-2006). In Figure 1, the dashed regression line in the middle denotes the fitted regression model for the entire sample that is assumed to come from a single, homogeneous population. However, the real data come from two different subpopulations with their own unique relationships between *X* and *Y* variables and those unique relationships are represented by two different straight lines, one in each ellipse of data. If the heterogeneous characteristics of the sample are ignored and a mixture modeling approach is not considered, the true relationship between the two variables will be distorted as in the dashed line.

Figure 1
An Example of a Regression Mixture Model



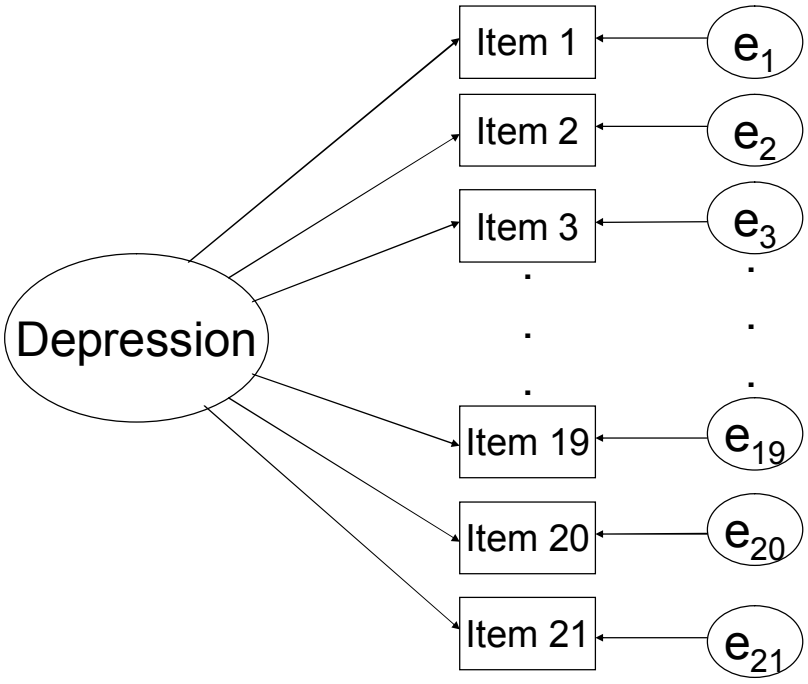
Along the same line, factor mixture modeling (FMM) delves into any present heterogeneity of factor structures in a single sample that may consist of heterogeneous subpopulations. FMM provides a flexible way of modeling factor structure heterogeneity in the data under investigation by combining confirmatory factor analysis (CFA) with latent class analysis (LCA), a special case of mixture modeling (Muthén & Muthén, 1998-2006). In FMM, CFA is used to specify a factor structure within each subpopulation and LCA is used to categorize and divide the entire sample into different subpopulations (Lubke & Muthén, 2005). FMM overcomes weaknesses of CFA such as the assumption of data homogeneity, by adding LCA into its framework, which allows for a more thorough investigation of construct validity of a psychological measure. To

better understand FMM, several relevant concepts and related traditional statistical methods are reviewed.

Latent Variables

In statistical analyses, variables can either be observed or latent. A latent variable can roughly be defined as an unobserved, hypothetical construct measured by a set of observed variables (Muthén, 2002). For example, the Beck Depression Inventory (BDI-II; Beck, Steer, & Brown, 1996) is hypothesized to assess levels of depression (i.e., a latent variable not directly observed) using scores on the 21 items (i.e., observed variables or indicators). These 21 items are considered to be manifestations of the underlying construct of depression and to represent different facets of depression (See Figure 2). The model described in Figure 2 is called a confirmatory factor analysis (CFA) model, where the latent construct (also called factor) of depression is measured by 21 indicators (i.e., items) or observed variables. The variance of each item consists of a portion that is due to the latent variable, depression, and a portion that is residual error (i.e., e_1, e_2, . . . , e_{21} in Figure 2) that is not explained by the factor.

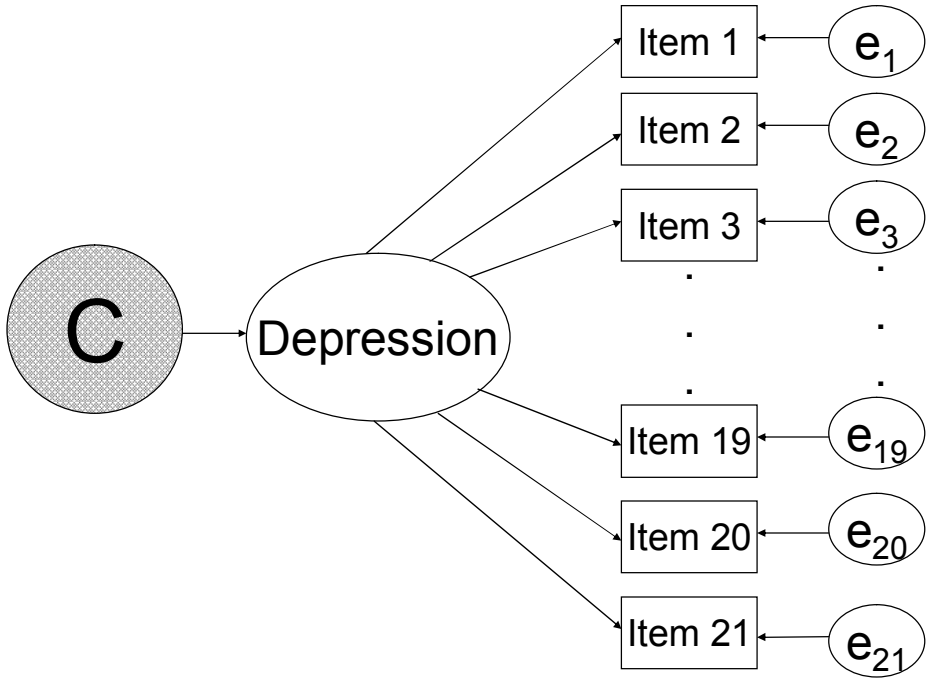
Figure 2
An Example of a Latent Variable Model: CFA



Similar to observed variables, latent variables can be either dependent or independent, and either categorical or continuous. In the BDI example shown in Figure 2, the latent variable, depression, could be considered an independent variable because it is conceptualized as an underlying cause of scores on its dependent indicators (i.e., Items 1-21). On the other hand, scores on the individual items can be considered dependent variables because they depend on the underlying construct of depression for their manifestation. In this dependent-independent relationship, it is assumed that the latent variable affects or causes the items (i.e., indicators) and that any score changes on the items of the BDI-II capture a change on the latent variable, depression (Borsboom, Mellenbergh, & Heerden, 2003). For example, the worse one's depression, the higher the scores on the observed BDI-II items.

It is also necessary to mention that a latent variable can be either continuous or categorical. In confirmatory factor analysis, factors are considered to be continuous. In the BDI example in Figure 2, the underlying construct of depression is presumed to lie on a continuum. On the other hand, latent variables can also be categorical (Bauer & Curran, 2004; Borsboom, Mellenbergh, & Heerden, 2003; Muthén, 2002). For example, we may be interested in dissimilar levels of depression across two latent classes. In Figure 3, these latent classes are represented by the C factor. Note that Figure 3 is simply an extension of a latent variable model depicted in Figure 2 with an added latent class variable, C, that is predicting levels of the Depression factor.

Figure 3
Factor Mixture Model: An Example Expanded on a CFA Model for Depression



Extending from the BDI depression example, these two latent classes could represent the *psychologically vulnerable* and the *psychologically invulnerable*. In line with factor analysis where factors (i.e., latent construct) are named, these latent groups or classes are typically named based on a substantive analysis that involves assessing what the members within each class have in common that distinguishes them from the members of other classes (Magidson & Vermunt, 2004). Note there is no interest here in differences between known, observed groups such as sex and ethnicity, but in classes that are inferred and not directly observed. Thus, in this example, the categories of *psychologically vulnerable* and *psychologically invulnerable* are latent (i.e., not readily observed).

Heterogeneity and Subpopulation

Before delving into factor mixture modeling, it is worthwhile for a better understanding of FMM to explain two essential relevant terms: heterogeneity and subpopulation. With most statistical analyses, it is usually assumed that a sample comes from a homogeneous population (Jedidi, Jagpal, & DeSarbo, 1997; Muthén, 1989; Yung, 1997), where all the individuals in the population come from the same distribution where the relationship among different variables are the same across all the individuals (Bauer & Curran, 2004). However, there are instances where this assumption of homogeneous populations cannot be held. A sample may consist of members of different subpopulations with different distributions (distinguished by parameter values or shapes). For example, it is well known that depression is more prevalent among women than among men (Benazzi, 2000; Hankin, Abramson, Moffitt, Silva, McGee, & Angell, 1998;

Kendler, Thornton, & Prescott, 2001). If a depression measure is administered among a random group of women and men, the mean score for women is much higher than men, reflecting the different prevalence rates. It is also probable that the standard deviations for each group will be different. In this case, data should be handled using an analytic technique that recognizes they come from heterogeneous populations or from a mixture of populations (Bartholomew, 2002; Muthén, 1989). If these two groups or subpopulations with different parameters (i.e., mean and standard deviation) are treated as coming from a homogeneous population, this would definitely distort the results of the statistical analysis.

With regard to this phenomenon, Lubke and Muthén (2005) defined subpopulation (e.g., men versus women in the above example) as a general term indicating a group within a heterogeneous population. However, heterogeneity can be observed or unobserved. When the membership of a subpopulation is unobserved, the subpopulation is called a class. For example, we can say that the depression rate will be higher among the psychologically vulnerable than among the psychologically invulnerable. However, the psychological vulnerability is not readily observed (in this example), but could be inferred based on, for example, people's responses to a questionnaire that measures the construct of psychological vulnerability. Also, it should be noted that heterogeneity is determined at the distribution level (Bauer & Curran, 2004). This suggests that a distributional moment (e.g., a mean or a variance) or any combination of distributional moments which uniquely define a distribution can differ across subpopulations. In statistics, the *k*th moment of a distribution is defined to be the

expected value of Y^k. Thus, the mean of a distribution becomes its first moment and the variance of a distribution is its second moment.

For example, subpopulations can be different only with respect to means, or only with regard to standard deviations, or with regard to a combination of both. Furthermore, as hypothesized in this study with the OQ-45 scores, different subpopulations may demonstrate unequal factor structures across classes, (related to the second moment of a distribution), even without factor mean differences (related to the first moment).

FMM mainly serves two purposes: (a) investigating varied factor structures and factor means across latent classes and (b) classifying groups of people without relying on observed grouping variables such as sex or race. FMM achieves these purposes by combining strengths of CFA (e.g., modeling a within-class factor structure in each class) and latent class analysis (e.g., assigning subjects into different latent classes) (Jedidi, Jagpal, & DeSarbo, 1997; Lubke & Muthén, 2005; Lubke & Muthén, 2007; Muthén, Asparouhov, & Rebollo, 2006). Both CFA and latent class analysis (LCA) utilize latent variables and try to explain covariation among a set of observed variables (Bauer & Curran, 2004; Borsboom, Mellenbergh, & Heerden, 2003). However, there are several other statistical methods also designed for the purposes of mean and factor structure comparison, as well as for classification (i.e., subject assignment). Even before FMM was introduced, these methods had already long been available. To better understand why to use FMM, as opposed to ANOVA, MANOVA and similar methods, it is essential to outline and critique the characteristics of these other more traditional statistical methods.

ANOVA and MANOVA

With regard to group mean comparison, various statistical methods are available. One important criterion for choosing an appropriate method is the kind of variable being compared, such as whether the variable is observed or latent. Researchers use ANOVA when comparing groups on a single observed dependent variable and MANOVA for a set of interrelated observed dependent variables (Hancock, 2004; Hancock, Lawrence, & Nevitt; 2000). For example, the severity of depression could be compared between two or more groups on Beck Depression Inventory (BDI-II; Beck, Steer, & Brown, 1996) scores using ANOVA. Alternatively, the groups could be compared on scores on both the BDI-II as well as the Hamilton Rating Scale for Depression (HRSD; Hamilton, 1960) using MANOVA.

MANOVA and ANOVA provide powerful tests used to examine group differences on a construct, however their limitations when compared with multi-group confirmatory factor analysis (MG-CFA; Jöreskog, 1971; Sörbom, 1974) have been extensively discussed (Bollen & Lennox, 1991; Cole, Maxwell, Arvey, & Salas, 1993; Hancock, 1997; Hancock, 2004; Hancock, Lawrence, & Nevitt, 2000). First, in the ANOVA framework, there is no modeling of measurement error. Hancock (2004) extensively discusses the unfavorable effect of measurement error on comparison of group means. In MANOVA, more weight tends to be assigned to the variable on which groups differ most. If the variable that most differentiates groups is not due to genuine difference in the construct being measured but due to the variable exhibiting more measurement error, MANOVA will still assign the most weight to that variable. As a

result, significant differences might be found, even though there might be no true group difference.

Second, in addition to the limitation of not taking measurement error into account, it has been suggested that MANOVA is only appropriate for a certain type of variable system, that is, an emergent variable system (Bollen & Lennox, 1991; Cole, Maxwell, Arvey, & Salas, 1993; Hancock, 1997; Hancock, 2004). In an emergent variable system, variables (i.e., indicators) influence a construct. As an example, consider a man who is diagnosed with major depression due to his job loss, divorce, and a car accident. Here, it is more reasonable to think that his depression (an emergent variable) is caused by his job loss, divorce, and a recent car accident than in an opposite direction, although the opposite is not impossible. On the contrary, in a latent variable system, indicators are assumed caused by an underlying latent construct. As another example, think about items on the BDI-II. Sleep loss, appetite change, and extensive crying are caused by depression (a latent variable in this case). A person loses sleep and appetite, and cries all day because he became depressed. It is not that the person became depressed due to his somatic and affective symptoms. The BDI-II items are effect indicators, not causal indicators as in an emergent variable system.

CFA and MG-CFA

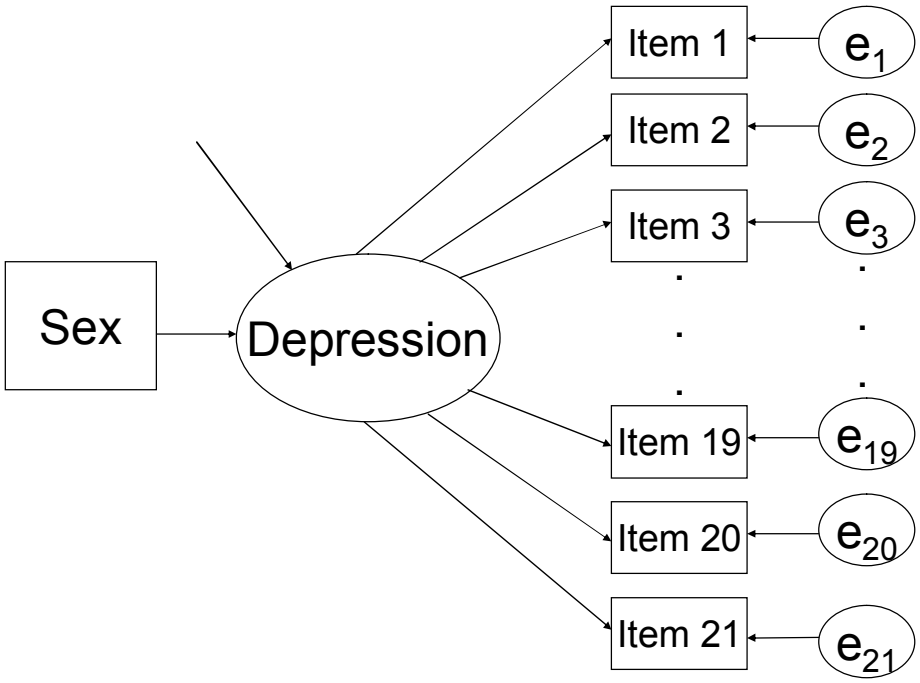
For a latent variable system, multi-group confirmatory factor analysis (MG-CFA) is recommended as a more appropriate method than MANOVA because it is based on a framework by which measurement error is modeled and controlled for (Hancock, 2004; Hancock, Lawrence, & Nevitt, 2000). CFA by itself does not allow one to compare

different groups due to its assumption that data come from a single population. However, Jöreskog (1971) and Sörbom (1974) extended the traditional CFA model and developed MG-CFA, where one can compare similarities and dissimilarities in factor structures and examine differences in factor means across observed groups. Since MG-CFA is based on CFA, a brief explanation of CFA is necessary.

CFA is a model that fits within the family of structural equation models (SEMs). Unlike exploratory factor analysis, while the number of factors and their intercorrelations is specified, the actual factor structure is not specified, CFA is used to assess an existing, pre-specified factor structure (Raykov & Marcoulides, 2006). CFA allows testing hypotheses about patterns of factor structure (Kaplan, 2000). In CFA, the covariance matrix implied by the model (represented as Σ) is tested against the covariance matrix produced by the data (represented as *S*). If there is a sufficient degree of fit between the proposed Σ and *S*, the proposed model can be regarded as a tenable representation of the observed relationships between latent variables and observed indicators (Kaplan, 2000; Raykov & Marcoulides, 2006).

As alternatives to MANOVA as a group mean comparison method, two MG-CFA methods have been suggested (Hancock, 1997; Hancock, 2004; Hancock, Lawrence, & Nevitt, 2000): Sörbom's (1974) structured means modeling (SMM) and multiple-indicator, multiple-cause (MIMIC) models (Jöreskog & Goldberg, 1975; Muthén, 1989). Figure 4 shows a MIMIC model for depression. Note that this MIMIC model is a simple extension of a CFA model (see Figure 2) with a covariate of sex. Also because the covariate is an observed variable, it is depicted within a rectangle, not within an ellipse as for a latent variable.

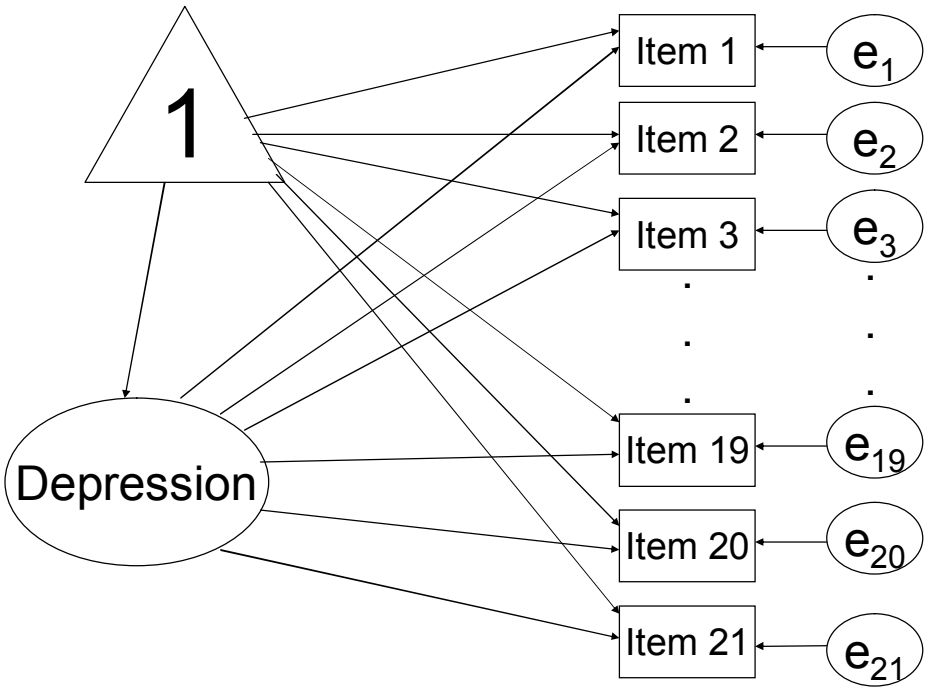
Figure 4
A MIMIC Model for Depression with a Covariate of Sex



A fundamental assumption of SEM and factor analysis is that the data under investigation come from a homogeneous population. For example, in Figure 4, although the factor model of depression is compared across different sex groups, the intercepts (i.e., the regression intercepts of 21 items on the depression factor), factor loadings, and residual variances are assumed to be equal between the sex groups. However, for the scenarios where there is heterogeneity in the form of differing sub-populations, Jöreskog (1971) demonstrated how a pattern of factor structure and loadings could be compared across different observed groups and suggested a strategy for evaluating similarities and differences of factor structures among observed groups with increasingly restrictive conditions.

Relying on Jöreskog's (1971) MG-CFA method, Sörbom (1974) devised a way of assessing factor mean differences across groups by adding intercepts into Jöreskog's (1971) model. Figure 5 shows an SMM model for depression. When different groups are compared in SMM, the samples from the different groups are separated before an analysis. Thus, unlike in a MIMIC model where data from different groups are combined into a single sample, an SMM model does not need a dummy covariate representing different groups such as the sex covariate in Figure 4. However, because SMM models the means of variables into its structure, it needs a unit-constant pseudo-variable, represented as 1 in Figure 5 (Hancock, 2004). In Figure 5 the paths (depicted with one-headed arrows) from the unit-constant to the items indicates the means of 21 items.

Figure 5
An SMM Model for Depression



SMM can be expressed as follows:

$$y_g = \nu_g + \Lambda_g \eta_g + \varepsilon_g \quad (1)$$

where y_g denotes a $p \times 1$ vector of observed indicators in group *g*; Λ_g is a $p \times 1$ vector of factor loadings; η_g denotes a latent variable, for example, reflecting the factor of depression in Figure 5; ν_g represents a $p \times 1$ vector of intercept values, depicted by arrows from the unit-constant 1 to 21 items in Figure 5; and ε_g is a $p \times 1$ vector of errors, also represented by e_1, e_2, . . . , e_{21} in Figure 5. The mean or first-order moment vector is expressed as $E(y_g) = \mu_g = \nu_g + \Lambda_g \kappa_g$, where κ denotes the mean of the latent variable η, or simply, the factor mean. Thus, SMM allows for mean comparison at the construct level testing the null hypothesis, $H_0 : \kappa_1 = \kappa_2$ for two groups. While MANOVA involves the comparison of only observed means, SMM decomposes the observed means into two components: the variable intercepts, ν and the latent variable means (Cole, Maxwell, Arvey, & Salas, 1993).

Close examination of the equation, $E(y_g) = \mu_g = \nu_g + \Lambda_g \kappa_g$ reveals that in order to compare the factor means across two groups (i.e., $\kappa_1 = \kappa_2$), the intercepts and the factor loadings should be the same across the groups under investigation (Hancock, Lawrence, & Nevitt, 2000). In other words, the observed variables should contain an equal amount of measurement bias (i.e., $\nu_1 = \nu_2$) and the measurement models of the groups compared should be tau-equivalent (i.e., $\Lambda_1 = \Lambda_2$), which requires a perfect correlation between true scores across groups (Allen & Yen, 1979; Raykov &

Marcoulides, 2006). This assumption of equal intercepts and factor loadings across groups is called "strong factorial invariance" (Meredith, 1993). The topic of measurement invariance and factorial invariance will be discussed in the following section.

In MIMIC modeling, on the other hand, strict factorial invariance is assumed. This means that the intercepts, factor loadings, factor variance, and error variances are expected to be equal across compared groups. In MIMIC modeling, because the heterogeneity between the two groups is in the factor means, the sub-populations' covariance matrices are assumed equivalent except for the covariance implied by the relationship between the group variable and the factor. Thus it is appropriate to treat the covariance matrices as equivalent. In the MIMIC model, factors are assumed to have means of zero and a dummy covariate representing the relevant groups being compared is included in the model to test for group differences on the factor mean (Hancock, 1997; Hancock, 2004; Hancock, Lawrence, & Nevitt, 2000). For example, in Figure 4 sex is a dummy coded covariate and represents males and females who are compared on their levels of depression that are measured with the BDI-II.

Although MG-CFA methods such as SMM and MIMIC models are superior to MANOVA for comparing groups' means on measures that lack perfect reliability, their limitation is that observed group membership must be known beforehand (Bauer & Curran, 2004). Thus, MG-CFA methods cannot be used when sources of heterogeneity are not readily observed or group membership is estimated in the process of model fitting (Bauer & Curran, 2004; Yung, 1997). Instead, FMM can be used where MG-CFA cannot be used because it investigates not readily observed, latent classes while still controlling for measurement errors in measured variables. FMM allows for dealing directly with

unobserved heterogeneity in factor structures and factor means because it is based on latent class analysis (LCA). However, before describing LCA, it is necessary to explore the concept of measurement invariance.

Measurement Invariance

Measurement invariance is defined as the absence of measurement bias with regard to group membership (Wicherts, Dolan, & Hessen, 2005). In other words, members of each group score the same on a test given they have the same ability so that there is no observed score differences between them. Meredith (1993) defined measurement invariance by drawing on conditional probability. For example, for a test item to be considered invariant across groups, the probability of endorsing that item should be equal among members of the same ability (i.e. conditional on ability) belonging to different groups, such as males and females (i.e., sex) and for Asian Americans and African Americans (i.e., race) (Lubke, Dolan, Kelderman, & Mellenbergh, 2003).

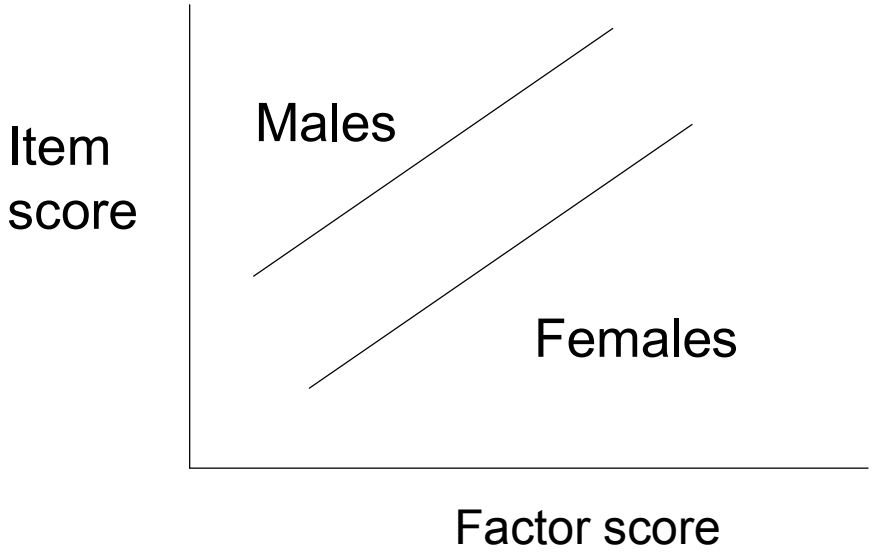
Figures 6, 7, and 8 (Lubke, Dolan, Kelderman, & Mellenbergh, 2003) depict three kinds of measurement bias caused by unequal intercepts, factor loadings, and residuals, respectively between two groups such as males and females. Figures 6, 7, and 8 show regression lines between observed scores on an item and the construct being measured like depression. The regression can be expressed as,

$$y = \nu + \Lambda \times \eta + \varepsilon \qquad (2)$$

If males' observed scores on the depression item are uniformly higher than females' although the two group's levels of depression are equal, it indicates the depression item

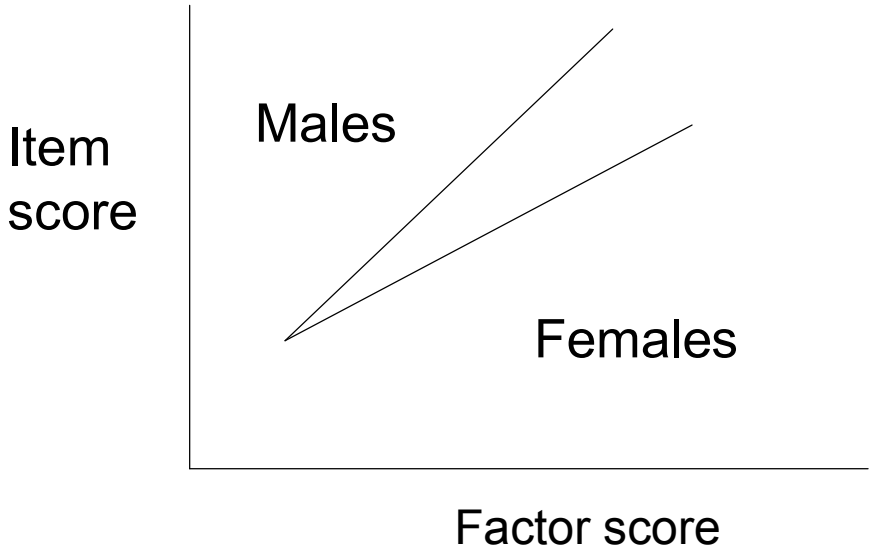
has a measurement bias (see Figure 6). This kind of measurement bias is called uniform bias (Lubke, Dolan, Kelderman, & Mellenbergh, 2003).

Figure 6
Measurement Bias Induced by Unequal Intercepts



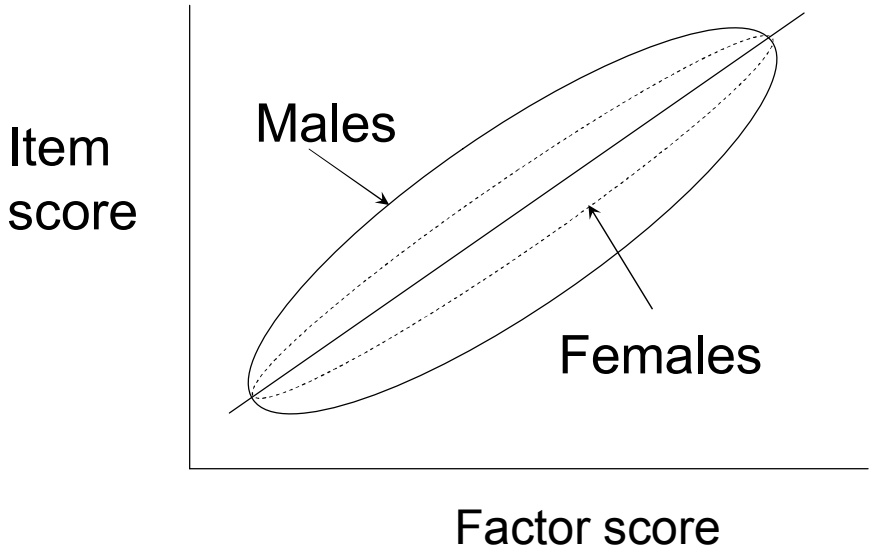
However, the measurement bias can be non-uniform across the levels of depression between groups. For example, a bias in the depression item between males and females can be wider among the more depressed than among the less depressed represented in Figure 7. This kind of a measurement bias reflects group differences in regression slopes (Figure 7) (Lubke, Dolan, Kelderman, & Mellenbergh, 2003). Note that the intercepts are the same across the two groups in Figure 7.

Figure 7
Measurement Bias Induced by Unequal Factor Loadings



Residual variances can also differ between groups thereby inducing another form of measurement bias. Figure 8 shows a case of measurement bias due to residual variance discrepancies across two groups. Note that, in Figure 8, the intercepts and the regression slopes are equal between the two groups so that the regression line is identical across males and females. However, the unequal residual variances depicted by the two ellipses of differing widths can introduce measurement bias. For example, because the residual variance of the depression item is greater among males, the probability of a false diagnosis for depression or non-depression is increased accordingly among males. On the contrary, the smaller residual variance for females implies less chance of a false diagnosis (Lubke, Dolan, Kelderman, & Mellenbergh, 2003).

Figure 8
Measurement Bias Induced by Unequal Residual Variances



Concurrent with the concept of measurement invariance, a different but closely related concept of factorial invariance, roughly defined as measurement invariance with respect to factor analysis, has been discussed (Allua, Beretvas, & Stapleton, 2006; Hancock, Stapleton, & Arnold-Berkovits, 2006; Meredith, 1993; Meredith & Horn, 2001; Widaman & Reise, 1997). Figures 6, 7, and 8 demonstrate how not only measurement invariance but also factorial invariance is violated by unequal intercepts, factor loadings, and residual variances in a simplified common factor model. The relationship portrayed with one item and one factor in Figures 6, 7, and 8 can be expanded in other relationships for several or more items and a few or more factors.

Meredith (1993) categorizes two kinds of factorial invariance: strict factorial invariance and strong factorial invariance. The condition of strict factorial invariance is

met when the intercepts, factor loadings, and residual variances in a factor model are equal across different groups. It means that none of the three kinds of measurement bias represented in Figures 6, 7, and 8 exist. However, note that these particular parts (i.e., intercepts, factor loadings, etc.) of invariance can also be combined. For example, only factor loadings and means can be non-invariant or, only intercepts and factor variance-covariance structures can be non-invariant between different groups.

In Equation 1, if y is measured in g different subpopulations (e.g., g different groups have taken the BDI-II), if the expected mean values of residuals (i.e., ε_g) are 0, and if y and the factors are not correlated, then the mean vector of y_g (e.g., scores of 21 BDI-II items across different groups being compared) is:

$$\mu_g = \nu + \kappa_g \Lambda \tag{3}$$

with a variance-covariance matrix for y, Σ_g:

$$\Sigma_g = \Lambda \Phi_g \Lambda' + \Psi \tag{4}$$

where κ_g and Φ_g denote the mean vector and variance-covariance matrix of latent factors in each subpopulation g. When Equations 3 and 4 are satisfied, y can be assumed to have strict factorial invariance (Meredith, 1993; Allua, Beretvas, & Stapleton, 2006; Lubke, Dolan, Kelderman, & Mellenbergh, 2003; Millsap & Yun-Tein, 2004). It should be noted that in Equations 3 and 4 the intercept, ν, the factor loading matrix, Λ, and the residual variance diagonal matrix, Ψ, do not have subscripts representing different subpopulations, representing the assumption that they are each equal across subpopulations. In other words, when strict factorial invariance is assumed,

subpopulations may be dissimilar only with respect to means (κ) and variance-covariances (Φ) of latent factors. Thus, any differences detected at the observed score level under the assumption of strict factorial invariance indicates that the subpopulations differ only in terms of their factor means and factor variance-covariance structures (Hessen, Dolan, & Wicherts, 2006; Meredith, 1993). That is why strict factorial invariance is strongly emphasized by some researchers, saying that observed scores cannot be compared properly unless strict factorial invariance holds (Lubke & Muthén, 2005). In another way, measurement invariance or factorial invariance is very important because any violation of it suggests that the factor structures (i.e., the measurement model) linking observed variables to factors are not identical across different groups and thus further comparison of factor structures or factor means across these groups will be futile (Lubke & Muthén, 2005).

When the condition of equal residual variances across groups is relaxed and they are allowed to vary (i.e., Ψ in Equation 4 becomes Ψ_g), y is regarded as having strong factorial invariance (Meredith, 1993; Meredith & Horn, 2001; Widamen & Reise, 1997). Thus, strict factorial invariance is a more restricted condition than strong factorial invariance. Under strong factorial invariance, only intercepts and factor loadings are assumed to be equal across groups. In addition to the conditions of strict and strong factorial invariance, Widaman and Reise (1997) discuss weak factorial invariance, a less restricted condition than strong factorial invariance.

Weak factorial invariance holds if only the factor loadings, Λ , can be assumed equal across groups without involving invariance of intercepts of measured variables and

the residual variances. These three different kinds of factorial invariance denote three different levels of equivalence of a measurement model between observed variables and latent factors across different subpopulations.

In the present study, strong factorial invariance where only intercepts and factor loadings are assumed to be equal across classes, will be deemed sufficient for a comparison of factor means and factor variance-covariances across different groups. There has been little consensus about the proper extent of factorial invariance that should be assumed for latent construct comparability (Lubke & Muthén, 2005; Allua, Beretvas, & Stapleton, 2006). According to Meredith (1993), strict factorial invariance is a necessary condition before factor means can be compared across subpopulations. On the other hand, Little (1997) argues that strong but not strict factorial invariance is a necessary condition when factor means are to be compared. He suggests that strict factorial invariance could potentially introduce a bias when measurement errors across different groups are not exactly equal. In line with Little, Widaman and Reise (1997) also argue that strong factorial invariance is enough to allow for meaningful comparison of latent variable means because any differences on the latent variables are appropriately reflected in differences on observed variables. In addition to strict and strong factorial invariance, weaker forms of factorial invariance have been discussed as a sufficient condition for latent construct comparisons across groups. Byrne, Shavelson, and Muthén (1989), for example, asserted that comparing factor mean differences can be pursued even under the condition of partial measurement invariance where neither all factor loadings nor all the intercepts can be assumed equal across groups. Therefore, there seems to be no consensus concerning what level of measurement invariance is required

for an appropriate comparison of latent mean differences (Allua, Beretvas, & Stapleton, 2006).

LCA (Latent Class Analysis)

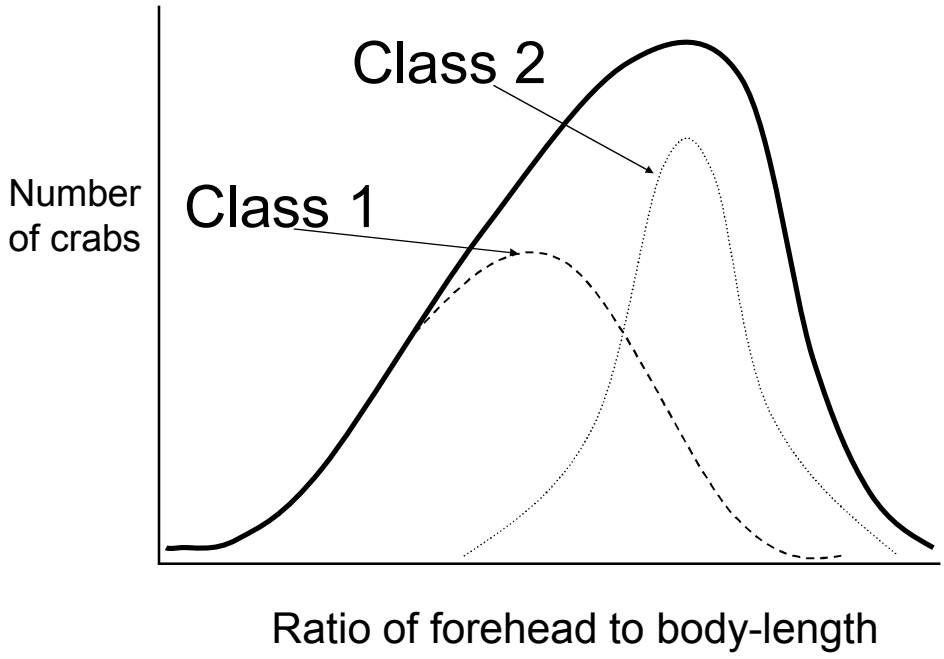
As mentioned above, FMM can handle unobserved heterogeneity in factor structures and factor means through combining LCA in its framework. It was also mentioned earlier that LCA is a special case of mixture modeling (Muthén & Muthén, 1998-2006). Although the framework of mixture modeling was introduced more than 100 years ago by the prominent British statistician Karl Pearson (1894, 1895), behavioral scientists have only recently come to pay more attention to this modeling procedure (McLachlan & Peel, 2000).

Pearson once was presented with a negatively skewed, univariate data set that measured the ratio of forehead to body length for 1,000 crabs from Naples. Faced with this kind of non-normal data, statisticians typically normalize the data as a strategy for facilitating statistical analyses. However, Pearson extracted two normal distributions with specific (and different) means and standard deviations from this single skewed distribution. Figure 9 represents the frequency table of the crab data, where the *x* axis represents measurements of the ratio of forehead to body length and *y* axis the number of crabs. Figure 9 shows one of the two mixture solutions Pearson derived. The outer solid line indicates the whole data set (when modeled as a single population) is negatively skewed and thus cannot be assumed normally distributed. The two inner dashed symmetric lines represent the two normal distributions (i.e., two subpopulations or classes) that Pearson fitted. Pearson stated the "asymmetry may arise from the fact that

the units grouped together in the measured material are not really homogeneous. It may happen that we have a mixture of 2, 3, … , *n* homogeneous groups…" (Pearson, 1894, p. 72). Pearson's normal mixture analysis supported the data donor's conjecture that the crab family was splitting into two different subspecies. Even though Pearson's analysis of the crab data was primarily concerned with univariate normal distributions, mixture modeling can be applied in more advanced way to samples with multivariate distributions (e.g., factor analysis) and latent constructs (Gagné, 2006).

Figure 9
The Single and Two-class Solutions for Pearson's Crab Data

Note. From "Contributions to the mathematical theory of evolution" by K. Pearson, 1894, *Philosophical Transactions of the Royal Society of London. A*, *185*, plate 1. Copyright 1894 by the Royal Society of London. Adapted with permission.



Mixture modeling has the primary utility of identifying a finite number of subpopulations whose members share similar response styles to measured variables

(Arminger, Stein, & Wittenberg, 1999; Greenbaum, Del Boca, Darkes, Wang, & Goldman, 2005; Lubke & Muthén, 2005; Nylund, Asparouhov, & Muthén, 2006). With regard to the primary aims of mixture modeling, Muthén (2002) also suggested additional goals including (a) modeling a (non-normal) mixture distribution and (b) investigating data heterogeneity with latent (i.e., unobserved) classes. Factor mixture modeling (FMM), a principal method to be used in the present study, falls in the family of mixture models (Yung, 1997).

LCA can be used to classify people into different categories (or classes) based on observed item responses (Nylund, Asparouhov, & Muthén, 2006). Figure 10 shows an LCA item profile, which depicts three hypothetical latent classes using observed item responses on six dichotomous items that measure somatic (i.e., Som1 and Som2), cognitive (i.e., Cog1 and Cog2), and affective (i.e., Aff1 and Aff2) symptoms of depression. Class 1 consists of a subpopulation whose endorsement probability on somatic and cognitive symptom items is much higher than on affective items. On the other hand, Class 2 is made up of a subpopulation whose symptoms are more cognitive and affective rather than somatic. Last, the members in Class 3 show low endorsement probabilities on all the items. As in factor analysis where names are designated to such factors as depression, anxiety, or psychological mindedness, names are also assigned to latent classes in LCA (Magidson & Vermunt, 2004). For example, we may designate Class 1 as a somatic-cognitive symptom class, Class 2 as a cognitive-affective symptom class, and Class 3 as a non-depressive or well-functioning class. In addition to classification, LCA aims to distinguish items that cluster people into different classes (Nylund, Asparouhov, & Muthén, 2006). The ability of LCA to cluster people into

different categories distinguishes it from CFA, which involves the assumption that sample data come from a homogeneous population (Bauer & Curran, 2004; Muthén & Asparouhov, 2006).

Figure 10
Latent Class Analysis: Item Profiles

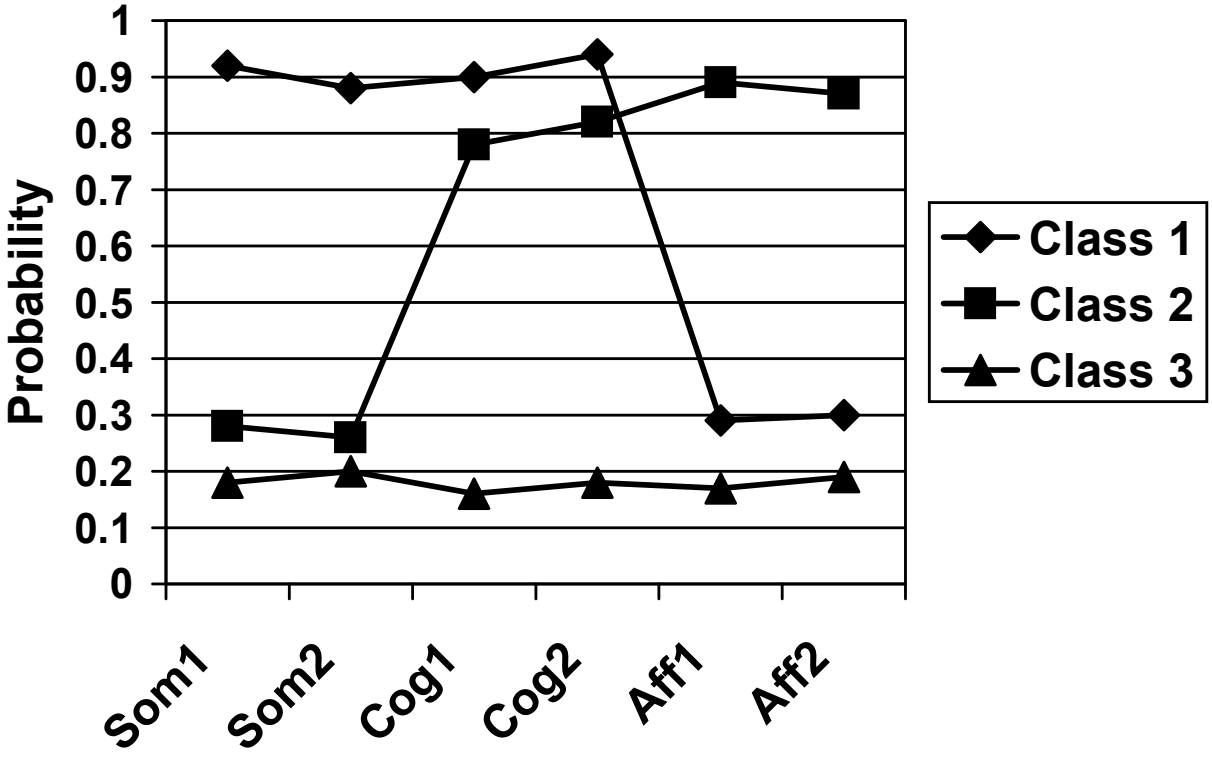
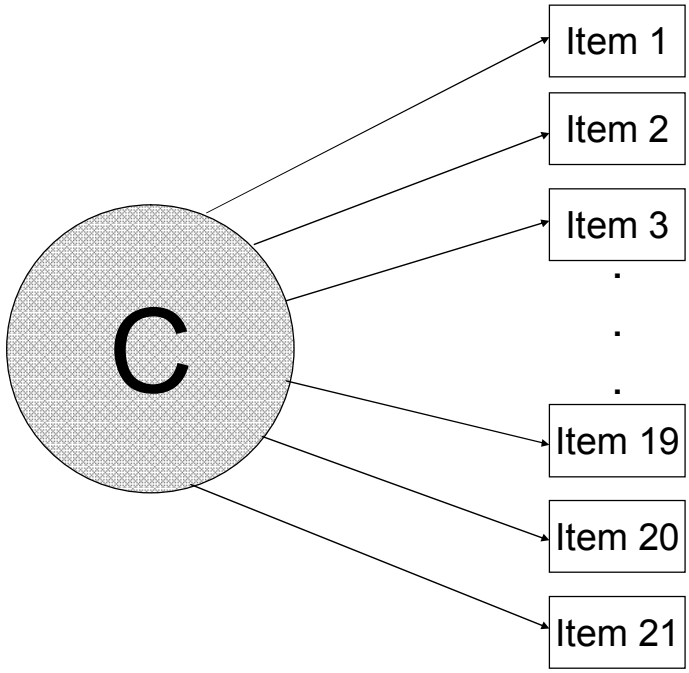


Figure 11 depicts a model diagram for LCA, where C in the shaded ellipse represents a categorical latent variable and items in rectangles indicate observed variables measuring various content domains of depression. Note the similarities and the differences between this figure and Figure 2, which contains a CFA model. The latent variable in Figure 2 is continuous, reflecting a common factor of depression. However, Figure 11 depicts a latent class variable (i.e., categorical) taking into account the correlation among the 21 items on the BDI-II. It should also be noted that in LCA, it is

assumed that observed variables (e.g., here, the 21 items measuring depression) are related to each other only through the latent categorical variable, C. Note that there are no error terms attached to individual items in Figure 11 in contrast to Figure 2. An objective of LCA is to find a minimum number of classes with which the maximum degree of covariation among observed variables can be explained (Magidson & Vermunt, 2004; Muthén, Asparouhov, & Rebollo, 2006).

Figure 11
Latent Class Analysis: Model Diagram



FMM: A Combination of CFA and LCA

It was stated that FMM aims to (a) investigate different factor structures and factor means across latent classes, and (b) assign subjects into different classes without relying on observed grouping variables such as sex or race. FMM achieves these

purposes by combining strengths of CFA (e.g., modeling a within-class factor structure in each class) and LCA, (e.g., assigning subjects into different latent classes) (Jedidi, Jagpal, & DeSarbo, 1997; Lubke & Muthén, 2005; Lubke & Muthén, 2007; Muthén, Asparouhov, & Rebollo, 2006). Both CFA and LCA utilize latent variables and try to explain covariation among a set of observed variables (Bauer & Curran, 2004; Borsboom, Mellenbergh, & Heerden, 2003). CFA and LCA employ different types of latent variables, namely continuous and categorical latent variables, respectively. In CFA the latent variable is known as a factor, which represents a common, continuous underlying dimension across a set of observed variables. In LCA, however, the latent variable is categorical in nature and represents discrete classes or subpopulations in a population. Although CFA and LCA use different types of latent variables, they achieve the same purpose of modeling factors designed to explain the common variance in a set of observed variables (Bauer & Curran, 2004; Borsboom, Mellenbergh, & Heerden, 2003).

Both LCA and CFA have their weaknesses. For example, LCA does not allow within-class covariation due to its assumption that covariances between observed variables (e.g., items) are all explained by a latent class variable (Muthén & Asparouhov, 2006; Nylund, Asparouhov, & Muthén, 2006). In LCA, if a model with a certain number (e.g., *N*) of classes does not fit data, it could be interpreted that the covariances among items are not accounted for with *N* classes; in this case, a model with $N + 1$ classes could be fitted.

FMM: Classification

Even though the number of classes is determined prior to the model estimation process, the membership of which class a subject belongs to is not known beforehand. The class membership is predicted during the estimation process and the probability of each subject's membership is calculated using multinomial regression (Lubke & Muthén, 2005). Although observed grouping variables such as sex and race may be included as covariate(s), they are not directly needed for classifying subjects into different classes in FMM. However, the covariate may help explain class membership (Lubke & Muthén, 2005). For example, one of the latent classes that were mentioned earlier, the *psychologically vulnerable* class, may have slightly more females and a high percentage of Whites when compared to the other class. On the other hand, the other class, the *psychologically invulnerable* class, might have slightly fewer males and more Hispanic Americans and Asian Americans.

The probability of a subject's falling into a specific class versus the other classes can be modeled as being influenced by the covariate x_i in a multinomial logistic regression equation (Muthén, 2004; Muthén & Shedden, 1999):

$$\ln\left[\frac{P(c_{ik}=1|x_i)}{P(c_{iK}=1|x_i)}\right] = \lambda_{ck} + \Gamma_{ck}x_i \qquad (5)$$

where, λ_{ck} denotes an intercept specific for each class and Γ_{ck} the regression weight for the covariate. For example, if there are two latent classes (c = 1, 2) representing *psychologically vulnerable* and *psychologically invulnerable* classes, respectively, and

the covariate indicates sex, with males coded with "1" and females coded with a "0", then Equation 5 can be specified as,

$$\ln\left[\frac{P(c_i = 1|x_i)}{P(c_i = 2|x_i)}\right] = \lambda_{01} + \Gamma_{11}x_i \qquad (6)$$

If $\Gamma_{11} = 1$, this indicates that the probability of belonging to the *psychologically vulnerable* versus the *psychologically invulnerable* class is almost three (about 2.72) times higher for males than females.

As discussed earlier, one of the general uses of mixture models such as LCA and FMM is to examine unobserved heterogeneity (Muthén, 2002; Nylund, Asparouhov, & Muthén, 2006). However, the issue of how to decide upon the appropriate number of classes or "class enumeration" in a given populations has not been resolved (Nylund, Asparouhov, & Muthén, 2006). Typically, in FMM, alternative models with an increasing number of classes are compared with each other to identify the best-fitting model with what seems to be the proper number of classes (Lubke & Muthén, 2005). For example, the null hypothesis and the alternative hypothesis can be expressed as,

$$H_0 : k = k_0 \qquad (7)$$

against

$$H_A : k = k_1 \qquad (8)$$

for some $k_1 > k_0$, where k denotes the smallest number of classes compatible with the data under investigation. In practice, the alternative hypothesis that is usually tested involves: $k_1 = k_0 + 1$. A χ^2 difference test cannot be used to test differences in fit of pairs

of mixture models which differ only in the numbers of classes being estimated (McLachlan & Peel, 2000).

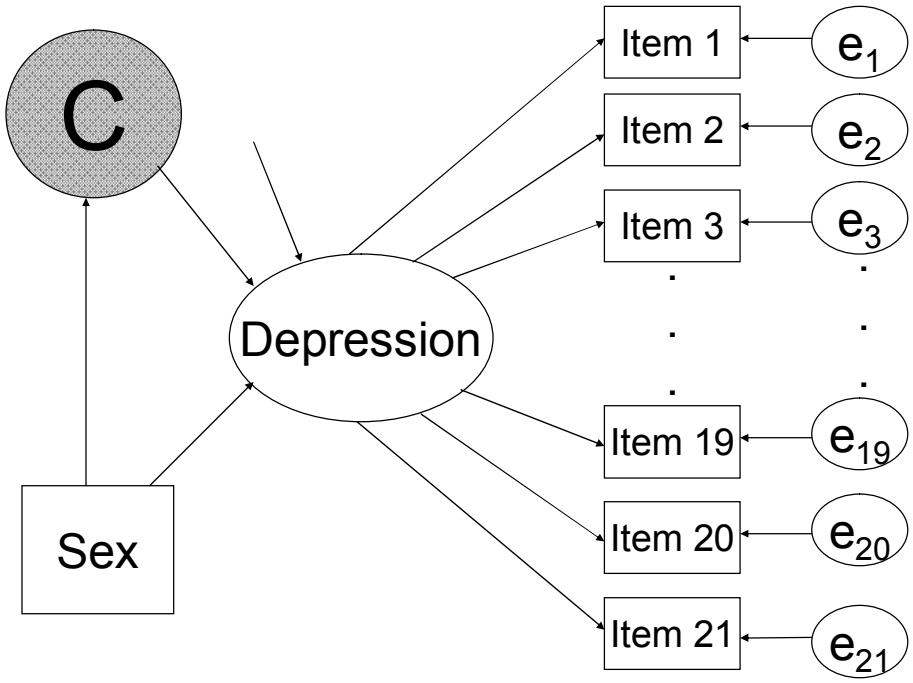
Because a χ^2 difference test cannot be employed to select the number of classes fitting a dataset, a combination of information criteria, such as Akaike Information Criterion (AIC; Akaike, 1987), Bayesian Information Criterion (BIC; Schwarz, 1978), and Adjusted BIC (ABIC; Sclove, 1987), are commonly used in the selection of the appropriate number of classes for a FMM (Jedidi, Jagpal, & DeSarbo, 1997; Muthén & Asparouhov, in press; Nylund, Asparouhov, & Muthén, 2006). These information criteria are calculated based on the log likelihood of a fitted model with a penalty for the number of parameters and/or sample size. A model with the lowest value for each information criteria among several fitted models is preferred over others (Lubke & Muthén, 2005; Nylund, Asparouhov, & Muthén, 2006).

In addition to these information criteria, Lo, Mendell, and Rubin (2001) proposed an adjusted likelihood ratio test (aLRT), a fit index calculated using an approximate distribution of the log likelihood ratio statistic. The aLRT provides a test, where a statistically significant *p*-value indicates the fit of the current model with *k* classes is better than the fit of the same model but with *k-1* classes. Also, the normalized entropy criterion (NEC; Celeux & Soromenho, 1996) is used to choose the number of classes and it indicates how well classes are separated. NEC ranges from 0 to 1 and larger values indicate better model fit (McLachlan & Peel, 2000; Muthén & Asparouhov, in press). However, when results of these criteria do not agree, and even when they do, relevant substantive theory should guide the model selection process (Gagné, 2006).

FMM: An Example and Equations

The notations and diagrams in this section follow the specifications by Muthén and Shedden (1999), and Lubke and Muthén (2005). Figure 12 depicts a factor mixture model with sex (i.e., males and females) as a covariate. Figure 12 is an extension of Figure 3 that shows a simple factor mixture model. The BDI-II example that was previously mentioned is used again to explain the features described in Figure 12. Suppose a sample is drawn from the population that is made up of two latent classes, namely, the psychologically vulnerable and the psychologically invulnerable. It is known that the two classes, represented by C vary markedly in terms of their depression levels. Note that the BDI-II is supposed to measure the factor of depression. Also, previous research (Benazzi, 2000; Hankin, Abramson, Moffitt, Silva, McGee, & Angell, 1998; Kendler, Thornton, & Prescott, 2001) has demonstrated that sex is related to the levels of depression. For example, it has been well supported that the prevalence of depression is much higher for women than for men. Thus, sex, a categorical observed variable is a known source of heterogeneity in the population. By including the covariate in the model more information about the composition of each class can be obtained.

Figure 12
An Example of a Factor Mixture Model with Sex as a Covariate



The FMM is expressed as

$$y_{ik} = \nu_k + \Lambda_{yk}\eta_{ik} + \Gamma_{yk}x_i + \varepsilon_{ik}\text{, and} \quad (9)$$

$$\eta_{ik} = Ac_i + \Gamma_{\eta k}x_i + \zeta_{ik} \quad (10)$$

where y_{ik} represents the observed dependent variables of subject *i* in latent class *k*. Note the similarity between this parameterization and that of MG-CFA where the *g* is the equivalent of the *k*, except that *g* indexes observed and *k* indexes unobserved groups. In Figure 12, for example, y_{ik} is a vector of observed scores for items 1 through 21. ν_k represents the regression intercept in latent class *k*. Parameters that have the subscript *k* may vary across classes. When strong or strict factorial invariance holds among latent classes this subscript is no longer needed for the relevant parameters. For example, the class-specific intercept, ν_k, can instead be assumed equal across classes and thus

represented as *v* if strict or strong factorial variance is assumed. Λ_{yk} denotes the matrix of factor loadings for class *k* and η_{ik} represents the matrix of factor scores. Covariate(s) are denoted by x_i for each subject *i* and the immediate effect of a covariate *x* on the outcome variable *y* is found in the matrix of regression weights, Γ_{yk}. The variable Sex (depicted in a box in Figure 12) represents a covariate. *A* denotes a matrix of intercepts for the factor scores in each class. c_i has a multinomial distribution, where $c_{ik} = 1$ when subject *i* belongs to class *k* and $c_{ik} = 0$, otherwise. In Equation 9, ε_{ik} indicates the residuals which consist of specific factor scores and measurement errors, and, in Equation 10, ζ_{ik} represents residual factor scores that are not accounted for by class membership and the covariate(s) in the model. Given the above example of an FMM model and the relevant equations, it is necessary to understand how subjects are classified into different classes and how an appropriate number of classes are determined in FMM.

FMM: Application and Growth Mixture Modeling

Despite its recognized usefulness, FMM has rarely been utilized in psychology except in very few applied studies, (e.g., Lubke and Muthén, 2005). However, growth mixture modeling (GMM; Muthén, 2002; Muthén & Shedden, 1999), a form of FMM, has been widely adopted in longitudinal research to investigate individual differences in growth trajectories over a period of time. In GMM, classes of individuals sharing a similar growth pattern are identified (Muthén, 2004). As in FMM, class membership of each individual is not observed, but the probability of class membership is estimated for each person (Muthén, 2004). GMM has been used to examine various issues, such as

juvenile delinquency (Reinecke, 2006; White, Bates, & Buyske, 2001), religious development in adulthood (McCullough, Enders, Brion, & Jain, 2005), alcohol consumption among adolescents and young adults (Colder, Campbell, Ruel, Richardson, & Flay, 2002; Greenbaum, Del Boca, Darkes, Wang, & Goldman, 2005; Jackson & Sher, 2005; Tucker, Orlando, & Ellickson, 2003), smoking among adolescents and young adults (Colder, Mehta, Balanda, Campbell, Mayhew, Stanton, Pentz, & Flay, 2001; Orlando, Tucker, Ellickson, & Klein, 2004), marijuana use from early adolescence to young adulthood (Ellickson, Martino, & Collins, 2004), and depressive symptoms in young men (Stoolmiller, Kim, & Capaldi, 2005). FMM also deserves a wide application in the behavioral sciences including psychology.

Clinical Implications of the Research

This study aims to uncover a proper factor structure of the OQ-45 using FMM. Once a sound factor structure is identified, it will affect clinical practices in several ways. For example, assume that the results of this study support a two-class (e.g., the psychologically vulnerable and invulnerable classes) factor mixture model, where a one-factor model of General Psychological Functioning is fitted within the psychologically invulnerable class but a two-factor model of Depression and Anxiety is fitted within the psychologically vulnerable class. Remember that FMM is a combination of LCA and CFA, where LCA categorizes individuals into different classes (i.e., the psychologically vulnerable and invulnerable classes) and CFA fits a factor model within each class by fitting the one factor model within the psychologically invulnerable class and the two-factor model within the psychologically vulnerable class (Lubke and Muthén, 2005). If

this is true, the results from previous research on the OQ-45 will not be applicable to the psychologically vulnerable class where the new factor model of Depression and Anxiety is fitted. For example, the cutoff score of 63 or 64 differentiating the clinical from the non-clinical populations are no longer meaningful because the score is based on the assumption that the OQ-45 consists of only one factor. For the individuals in the vulnerable class where the two-factor model is fitted, two cutoff scores will be necessary, one each for Depression and Anxiety. For the same reason, the reliable change index (RCI) of 14 points will need to be modified and two RCIs for the two factors will be required for an effective use of this index. However, prior to applying these new cutoff scores and RCIs, it is essential the therapist first find out which class the client belongs to. Since classes are unobservable in FMM, observed traits or covariates in the model should be used to help the class membership identification process. In this study, each individual's sex, race, and clinical status will be included as covariates. For example, the resulted model may specify the probability of males belonging to the psychologically invulnerable class is .9, but that of females belonging to the same class is .1. If a client is a male, it would then be more likely that he is a member of the psychologically invulnerable class, where the one-factor model was fitted. However, the final decision on his class membership should be made considering more information about other background variables, for example, his race and clinical status. Although, after the identification of a new factor structure, all the ensuing score interpretation and clinical decision-making procedures sound complicated, scores on the OQ-45 that has undergone a more rigorous validation process will become more informative and effective outcome assessment tools. Furthermore, if this new factor mixture model is proved to be

appropriate, the resulting, new system of score interpretation will inevitably affect the practices and policies (e.g., cost reimbursement and service authorization for patients) of managed health care organizations that rely on the LSQ, 30-item version of the OQ-45, as a therapy monitoring assessment tool.

Statement of Purpose

The OQ-45 is very unique in terms that it is specifically designed for repeated measurement of psychotherapy outcome through sessions up to termination in a wide variety of clinical settings (Lambert et al., 1996). Also, it taps into various domains of psychological functioning among adults and can be utilized as a baseline assessment tool (Lambert, Gregersen, & Burlingame, 2004). Its ease of administration and scoring, and low cost makes itself an attractive choice of outcome measurement (Lambert et al., 1996). Therefore, it is not surprising that it has been so eagerly adopted by both practitioners including a managed health care organization and researchers alike that it has become one of the most extensively utilized outcome monitoring measures (Hatfield & Ogles, 2004).

Despite its broad usage and future potential in both clinical and research settings, unsatisfactory results have been found with the OQ-45 with respect to its psychometric functioning. First of all, no study has resulted in decent support for the factor structure of OQ-45 scores. The OQ-45 is deemed to consist of either one general factor reflecting psychological distress or three factors representing the three subscales of the OQ-45. However, the only two studies available on the factor structure of the OQ-45 have failed to support either of those factor models (Beretvas & Kearney, 2003; Mueller, Lambert, &

Burlingame 1998). Second, studies comparing different sexes and races on OQ-45 mean scores have not yielded consistent results. This implies problems related to measurement bias and validity concerns and indicates a possible need for separate norms. ANOVA that has principally been used to compare groups on the OQ-45 score means has been inadequate to deal with these issues. So far, in the research on the psychometric functioning of the OQ-45, these problems have not been given serious attention in spite of their potential significance in both clinical practice and research.

Factor mixture modeling (FMM), a modeling method that addresses these limitations of previous research, is utilized in the present study to investigate similarities and differences in factor structures and factor means across different classes. While FMM has been used, it seems that it is not sufficiently used in the field of psychology and, in particular, in the field of validation of psychological test scores. FMM provides a flexible alternative to traditional statistical methods such as CFA and ANOVA by incorporating both latent continuous and categorical variables in its framework. Also, FMM allows exploration of unobserved sources of possible heterogeneity.

Chapter 3: Methodology

PURPOSES OF THE STUDY

As mentioned in the previous chapter, the current study aims to address the weaknesses of previous OQ-45 validity studies by using FMM to explore potential heterogeneity in scores on the OQ-45. Specifically, this study has three different, but closely related, purposes.

First, it is designed to identify an appropriate factor structure for OQ-45 scores. For this purpose, guided by theoretical premises, two confirmatory factor models (i.e., one modeling one overall factor and the second modeling three correlated factors) will be fitted to OQ-45 scores. However, if, as previous research demonstrated, neither model fits the data well, an exploratory factor analysis (EFA) will be attempted to identify a decent factor model.

Second, the present research aims to assess whether there is any heterogeneity in the factor structure across unobserved subpopulations. Here, factor mixture modeling will be employed to investigate any discrepancies with regard to factor loadings, intercepts, and errors in factor structure among classes. The two CFA models (the one-factor and three-factor models) as well as the potentially new factor structure identified with the EFA in the previous step will each be fitted as a within-class structure in two-, three-, and four-class factor mixture models. Two-, three-, and four-class versions of the factor mixture models will be considered due to model parsimony and interpretability. In FMM the researcher should specify the number of classes prior to data analysis. Because factors

of the OQ-45 items are assumed to measure various dimensions of psychological functioning, it is likely that latent classes will be distinguished by levels on those same dimensions. Thus, it is hypothesized that there will be differences in OQ-45 factor means across the latent classes.

However, there should be a sufficient degree of factorial invariance across classes in order for factor means to be compared. As stated before, there has been little agreement on the appropriate scope of factorial invariance that should be assumed for latent construct comparability (Lubke & Muthén, 2005; Allua, Beretvas, & Stapleton, 2006). In the present study, strong factorial invariance where only intercepts and factor loadings are assumed to be equal across classes, will be deemed sufficient for factor mean comparison. Although strict factorial invariance is an optimal precondition for latent mean comparison, it is known that strict factorial invariance induces a bias when measurement errors are not exactly equal among different classes (Little, 1997). Also, it has been argued that strong factorial invariance is a sufficient condition for meaningful comparison of latent variable means because any differences on the latent variables are appropriately reflected in differences in observed variables (Widaman & Reise, 1997).

Third, the present study attempts to investigate the relationship between subjects' observed characteristics that may introduce heterogeneity and their latent class membership by incorporating three covariates, specifically: sex, race, and clinical status (i.e., a clinical group and a non-clinical group) into the factor mixture model supported in the previous step. Note that other race groups other than Whites are combined into a single race group of non-Whites because Whites make up majority (80.9%) of the total

sample and the sample sizes of the other races seem too small for stable parameter estimation.

DATA DESCRIPTION

The Research Consortium of Counseling and Psychological Services in Higher Education (the Research Consortium or RC) was established in 1990 to investigate the effectiveness of counseling services for college students' mental health across nationwide. RC data have been collected for four cohorts since 1991; in this study, data sets from the third and fourth cohorts are combined and analyzed. The RC Project 3 (i.e., the project for the third cohort) was concerned with the psychotherapy process/outcome of clinical populations in colleges who came to college counseling centers for psychological help. In this project students were recruited during the 1997-98 school years. Students who agreed to participate in the project completed a consent form and were given several questionnaires prior to intake, which assessed clients' counseling concerns, presenting problems, working alliance, and preparation for change. Clients also completed the OQ-45 prior to the intake and also prior to the subsequent individual sessions. In Project 3, data were collected on 4,679 clients across 42 college counseling centers nationwide.

Data collection for the next cohort (i.e., Project 4) was resumed four years later during the 2001-2002 school years as a part of the RC's ongoing effort to create a database which encompasses both a clinical population and a non-clinical population in colleges. For the fourth cohort, a non-clinical subpopulation that was not seeking professional psychological help at the time of survey was recruited. This sample was to

be compared with a clinical sample to examine the differences in psychological concerns between both populations. Students filled out the same measures used in Project 3. However, they were not assessed using a measure of working alliance. In addition, they completed the OQ-45 only once. Data on 1,586 students were obtained across 16 counseling centers for this Project 4 cohort.

For the present study, the two data sets from Projects 3 and 4 are combined. Also, only OQ-45 scores at the initial survey from both projects will be analyzed because these are the sole OQ-45 data point that was measured at comparable time points across the two cohorts. Because sex and race have been indicated as possible sources of group differences on OQ-45 total and subscale scores in previous research and, in addition, the diagnosis (i.e., whether a subject is from Project 3, a clinical group or Project 4, a non-clinical group) is hypothesized to introduce heterogeneity in the data, these three variables will be included as covariates in the FMMs. An FMM analysis requires that all the cases that are missing any covariate values be removed from the analysis. Hence, 4,484 cases were selected out of 6,265 participants in the combined data set after deleting all the cases that are missing any covariate values. In addition, 24 cases who responded to fewer than 35 items (78% of the OQ-45 items), were deleted from the data. Furthermore, 221 international subjects were removed from the dataset. In the end, 4,239 cases were identified for analysis consisting of 2,874 cases (67.8%) from the third cohort and 1,365 cases (32.2%) from the fourth cohort.

The final sample included 2,458 (58.0%) men and 1,781 (42.0%) women. The average age was 22.4 years (SD = 5.21). The race breakdown for the sample was as follows: 3,428 (80.9%) Whites, 399 (9.4%) Hispanic-Americans, 230 (5.4%) Asian-

Americans, 169 (4.0%) African-Americans, and 13 (0.3%) Native Americans. The number of non-Whites was 811 (19.1%). The majority (n = 3,566, 84.5%) of the sample was enrolled in undergraduate programs. Table 2 describes the final sample in terms of the three covariates: race, sex, and clinical status.

Table 2
Composition of the Total Sample (Race, Sex, and Clinical Status)

Race	*n* (%)	Sex	*n* (%)	Clinical Status	*n* (%)
White	3,428 (80.9%)	Male	2,458 (58.0%)	Project 3	2,874 (67.8%)
Non-White	811 (19.1%)	Female	1,781 (42.0%)	Project 4	1,365 (32.2%)
Hispanic[a]	399 (9.4%)				
Asian[b]	230 (5.4%)				
African[c]	169 (4.0%)				
Native[d]	13 (0.3%)				
Total	4,239 (100%)		4,239 (100%)		4,239 (100%)

Note. [a]Hispanic-American. [b]Asian-American. [c]African-American. [d]Native American.

DATA ANALYSIS

Data analysis consisted of three major steps and each step required its own exclusive randomly selected sub-dataset. The three steps included estimation of the following models: (a) CFA and EFA, (b) FMMs for the second and third purposes in the study, and (c) a cross-validation of the final fitted model using a new random sub-sample. Thus, it was necessary to randomly select three evenly divided sub-datasets from the entire, final sample of 4,239 cases. However, given that Whites and the clinical population (i.e., Project 3 cohort) constitute more than three fourths (80.9%) and two thirds (67.8%) of the total sample respectively, stratified random sampling needed to be carried out after dividing the entire sample into four distinctive groups (i.e., White

clinical, White non-clinical, non-White clinical, and non-White non-clinical). Then, each distinctive group was randomly divided into three sub-groups and each sub-group was combined with the other three corresponding sub-groups. Finally, three unique sub-datasets for the three major steps of this study were obtained and named Sample 1, Sample 2, and Sample 3, respectively. The composition of these three sub-datasets is very similar to the total sample in terms of the three covariates: race, sex, and clinical status (see Table 2). The number of subjects in each sub-dataset is identical (n = 1,413).

For the initial CFA, a one-factor model and a three-factor model was fitted. CFA was conducted using maximum likelihood (ML) estimation procedure in Mplus version 4.1 (Muthén & Muthén, 2006). Fit indices employed for a CFA model evaluation included the model χ^2, the Comparative Fit Index (CFI), the Non-Normed Fit Index (NNFI), Standardized Root Mean Square Residual (SRMR), and the root mean square error of approximation (RMSEA). A value of .95 or greater was used to reflect a good model fit for the CFI and NNFI (Hu & Bentler, 1999; Raykov & Marcoulides, 2006). A value of .08 or less is recommended as a good fit for SRMR and a value of .06 or less was used to represent a good model fit for RMSEA (Hu & Bentler, 1999).

However, because the two CFA models were not proven to be adequate for the data, it was proposed another possible model be explored using not only modification indices but also an exploratory factor analysis (EFA). In an EFA, the number of factors to be extracted is determined by using a combination of the eigenvalues, the scree plot, relevant theories in psychology and mental health informing an item content analysis. In this EFA, a loading's magnitude of .40 or larger was considered substantial factor loading and items with substantial factor loadings were retained (Hair, Anderson, Tatham, &

Black, 1998). EFA was carried out using SPSS 13.0. In particular, principal axis factoring was used as the extraction method with the extracted factors rotated using the Direct Oblimin oblique rotation.

For the second and third purposes, FMM was utilized. The factor models with one overall factor and three sub-factors as well as a new factor model that was identified with modification indices and an EFA served as baseline models with which subsequent factor mixture models can be compared and evaluated. Note that these baseline models are not mixture models because each of these was estimated with one class specified. It was expected that fitting the baseline models would result in worse fit than fitting corresponding factor mixture models with two-, three-, and four-classes. A combination of fit indices, including the AIC (Akaike, 1987), BIC (Schwarz, 1978), aBIC (Muthén & Muthén, 2006), and aLRT (Lo, Mendell, and Rubin, 2001), was used to compare models with different number of classes. If, contrary to expectation, the baseline, one-class model fitted better to the data than any other models with more classes, the present study would simply become a CFA study. Fit indices such as CFI, NNFI, RMSEA, and SRMR were used for a CFA model evaluation.

As the final step, the best fitting model was selected as the final model after all these different models were compared. However, the same set of models that had been fitted in the second step were cross-validated using the third randomly selected subset of data in hopes that support can be found for the same final model that had been identified in the second step. Factor mixture modeling estimation also was conducted using Mplus version 4.1 (Muthén & Muthén, 2006).

Chapter 4: Results

CONFIRMATORY FACTOR ANALYSIS

Confirmatory factor analyses were conducted with maximum likelihood estimation in order to examine the factor structure of scores on OQ-45 items for Sample 1, a randomly selected subsample of 1,413 participants. The fit of the one- and three-correlated factor models were investigated. All of the 45 items were included in the one-factor CFA model, but only 42 items (i.e., all but items 11, 27, and 40) were included in the three-factor model. Note that the three subscales of the OQ-45, Symptom Distress, Interpersonal Relations, and Social Role Performance consist of 22, 11, and 9 items respectively, and do not use items 11, 27, and 40 (Beretvas & Kearney, 2003; Mueller, Lambert, & Burlingame, 1998). Nevertheless, the scores on these three items are included in the calculation of the total score, which represents the single factor in the one-factor model.

The one-factor model did not provide a good fit to the data (χ^2 (945, n = 1,413) = 8,926.88, p < .0001, CFI = .697, TLI = .683, RMSEA = .077, SRMR = .068). The values of all the fit indices did not meet the recommended criteria (see Table 3), indicating a poor fit of the model to the data. Although the fit of the three-factor model appeared better than that of the one-factor model, the model fit indices still did not support model fit (χ^2 (814, n = 1,413) = 7,231.72, p < .0001, CFI = .737, TLI = .722, RMSEA = .075, SRMR = .066).

Table 3
Goodness-of-Fit Indices for the One- and Three-Factor Models

Model	χ^2	*df*	χ^2 / *df*	CFI	TLI	RMSEA	90% CI[a]	SRMR
One-factor	8,926.88	945	9.45	.697	.683	.077	.076 - .079	.068
Three-factor	7,231.72	814	8.88	.737	.722	.075	.073 - .076	.066

Note. n = 1,413. CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual. [a] 90% confidence interval for RMSEA.

The factor loadings, the average proportion of variance (R^2) accounted in each item by the latent factors in the one- and three-factor solutions, and modification indices were investigated to find sources (i.e., items) of model misfit. All the items except Item 14 in both factor models had significant factor loadings. However, the factor loadings on 11 items in the one-factor model and five items in the three-factor model were less than .40. Furthermore, the modification indices (MIs) were examined for a source of model misspecification. The MI represents the value of the expected drop in the model chi-square if the corresponding fixed parameter were freely estimated. Also, associated with each value of MI is an expected parameter change (EPC) index. An EPC index denotes the expected value for a fixed parameter if it were freely estimated (e.g., Byrne, 2001; Muthén, L. K. & Muthén, B. O., 1998-2006).

In the three-factor model, the values of seven MIs and their associated standardized EPC indices were substantial (i.e., greater than .40). The largest standardized EPC index was found for Item 22 on the Social Role Performance factor with a value of .93. This means that if Item 22, which is designed to measure Symptom

Distress, is allowed to load on the Social Role Performance factor its standardized regression coefficient would be expected to be around .93. Other large values were identified for Item 21 and Item 38 standardized EPC indices of .88 and .53, respectively. All of these findings indicated that neither the one- nor the three-factor models proposed by the developers of the OQ-45 fit the data. Given this inadequate model fit for both the one- and three-factor models, it was decided to pursue an exploratory factor analysis (EFA) to identify a plausible alternative factor structure for scores on OQ-45 items.

Exploratory Factor Analysis

An EFA sometimes precedes and facilitates a following CFA in an exploration of a new factor structure. The rationale of this two-step approach was advocated in Gerbing and Hamilton (1996), who recommended conducting an EFA to help researchers identify a factor structure that can be verified subsequently with CFA. In this section, first a new factor model of scores on OQ-45 scores will be explored with an EFA and then the new model will be validated using CFA techniques.

As the initial step to select the number of factors to be extracted, the 45 items were content-analyzed. The result of the content analysis suggested four to seven factors (e.g., Depression, Relationship Satisfaction, Social Role Functioning, Somatization, Sense of Well-being, Substance Abuse, and Anxiety) with some items possibly cross-loading on different factors. Next, an exploratory factor analysis was conducted using principal axis factoring and Direct Oblimin oblique rotation. Common factor models of four-, five-, six- and seven-factor solutions were compared and examined. Direct Oblimin oblique rotation method was chosen because the factors were expected to be correlated.

The delta weight was set to zero, allowing a moderate correlation among factors. The factor eigenvalues and the scree plot were examined to assist in selecting an appropriate number of factors. Only the items with factor loadings greater than or equal to .40 on a primary factor and less than or equal to .30 on non-primary factors were kept in the factor solution.

Following a close examination of these four-, five-, six- and seven-factor solutions, the four-factor solution was finally chosen over the other solutions for the following reasons. First, the scree plot supported the five factor solution. Nevertheless, it was revealed that the fifth factor in the five-factor solution only had two items available, thus not meeting the criterion of at least three items per factor for reliable measurement (Floyd & Widman, 1995). Second, the six- and seven-factor solutions were examined. As with the five-factor solution, one or more factors from each solution would be deleted because the number of items that meet the substantial factor loading criterion was less than three (Floyd & Widaman, 1995). Finally, the contents of the items in the four-factor solution were closely examined to see whether the four-factor solution is justified from a substantive standpoint. The item contents and the factor loadings of the four-factor solution are presented in Table 4. The four-factor inter-correlation matrix is displayed in Table 5. Clearly, most of the items that did not meet the substantial factor loading criteria also did not measure the factors they were supposed to. For example, the nine items that did not meet the substantial factor loading criteria in the first factor, except Items 34 and 35, appear to measure something other (e.g., self-blame, family trouble, unwanted thoughts, arguments, etc.) than what the other items measure (see the first pane of Table 4). In addition, most of the six items in the second factor that do not meet the factor

loading criteria seem to measure something distinctive (e.g., suicidal ideation and unfulfilling sex life) from what the other items measure. This phenomenon also can be seen with the other two factors.

Table 4
Item Factor Loadings for an Exploratory Factor Analysis Solution of Four-factor Model

Factor and item	Factor 1	2	3	4
Factor 1: Anxiety-Somatization				
36. I feel nervous.	**.568**	-.071	-.019	-.075
45. I have headaches.	**.548**	.111	-.021	-.181
27. I have an upset stomach.	**.515**	.046	.028	-.099
33. I feel that something bad is going to happen.	**.508**	-.155	.067	-.063
29. My heart pounds too much.	**.504**	-.038	.023	-.098
10. I feel fearful.	**.475**	-.195	-.060	-.097
9. I feel weak.	**.437**	-.215	.017	-.246
41. I have trouble falling asleep or staying asleep.	**.402**	-.145	.038	-.130
34. I have sore muscles.	.385	.037	.044	-.045
25. Disturbing thoughts come into my mind that I can't get rid of.	.381	-.195	.136	-.061
6. I feel irritated.	.346	-.246	.081	-.088
5. I blame myself for things.	.336	-.244	.040	-.210
40. I feel something is wrong with my mind.	.328	-.286	.139	-.177
19. I have frequent arguments.	.313	-.192	.165	.067
16. I am concerned about family troubles.	.305	-.108	.064	.037
35. I feel afraid of open spaces, or of driving, or being on buses, subways, etc.	.283	-.035	.080	.011
14. I work/study too much.	.278	.151	-.084	.160
Factor 2: Negative Self-worth				
43. I am satisfied with my relationships with others.	-.001	**-.718**	-.061	.004
20. I feel loved and wanted.	-.041	**-.704**	-.014	-.019
24. I like myself.	-.049	**-.645**	-.014	-.152
13. I am a happy person.	.002	**-.629**	-.043	-.227
31. I am satisfied with my life.	-.004	**-.617**	-.035	-.196
37. I feel my love relationships are full and complete.	-.040	**-.582**	-.005	.030
18. I feel lonely.	.188	**-.578**	-.044	-.052

(table continues)

Table 4 (continued)

Factor and item	Factor 1	2	3	4
1. I get along with others.	-.088	**-.490**	.042	.041
21. I enjoy my spare time.	.029	**-.445**	-.070	-.201
15. I feel worthless.	.214	**-.430**	.091	-.177
42. I feel blue.	.317	-.429	-.070	-.238
30. I have trouble getting along with friends and close acquaintances.	.222	-.367	.133	.059
23. I feel hopeless about the future.	.183	-.360	.104	-.307
8. I have thoughts of ending my life.	.173	-.338	.240	-.025
17. I have an unfulfilling sex life.	.101	-.329	.065	.163
7. I feel unhappy in my marriage/significant relationship.	.215	-.288	-.028	.096
Factor 3: Substance Abuse				
26. I feel annoyed by people who criticize my drinking or drug use.	-.044	.105	**.639**	.017
32. I have trouble at work/school because of drinking or drug use.	-.070	.085	**.567**	-.085
11. After heavy drinking, I need a drink the next morning to get going.	.006	.054	**.492**	.072
39. I have too many disagreements at work/school.	.112	-.092	.308	-.068
44. I feel angry enough at work/school to do something I might regret.	.232	-.143	.291	-.068
Factor 4: Loss of Interest				
22. I have difficulty concentrating.	.218	.047	.081	**-.596**
28. I am working/studying less well than I used to.	.155	.038	.044	**-.556**
38. I feel that I am not doing well at work/school.	.068	.013	.101	**-.546**
12. I find my work/school satisfying.	-.113	-.223	.036	**-.508**
4. I feel stressed at work/school.	.317	-.013	-.063	-.418
3. I feel no interest in things.	.116	-.265	.070	**-.411**
2. I tire quickly.	.345	-.029	-.047	-.374

Note. Those factor loadings that meet the substantial factor loading criterion are in bold face and underlined.

Table 5
Factor Inter-correlation Matrix: Four-factor Model

Factor	1	2	3
2	-.399		
3	.221	-.274	
4	-.316	.386	-.193

The four-factor model consisted of 26 items. Together, the four factors accounted for 39.88% of the total variance. The first factor was made up of eight items (items 9, 10, 27, 29, 33, 36, 41, and 45) that mostly measure anxiety and somatization and named Anxiety-Somatization (e.g., "I have an upset stomach"). 10 items (items 1, 13, 15, 18, 20, 21, 24, 31, 37, and 43) comprised the second factor and tap into a negative sense of worth in one's interpersonal relationships and the self. The second factor was named Negative Self-worth (e.g., "I feel worthless"). The third factor consisted of three items (items 11, 26, and 32) related to drug and alcohol use in one's life and was named Substance Use (e.g., "I feel annoyed by people who criticize my drinking (or drug use)"). The fourth factor consisted of five items (items 3, 12, 22, 28, and 38) that mostly evaluate one's level of interest at work and school and thus was named Loss of Interest (e.g., "I feel that I am not doing well at work/school").

Confirmatory Factor Analysis on the Four-Factor Model

A CFA was then conducted to validate the newly identified factor model using Sample 2 (n = 1,413). The fit of the four-factor model was much improved over that of the one- and three-factor models, but still was not found to fit the data (χ^2 (293, n = 1,413) = 2,115.43, $p < .0001$, CFI = .872, TLI = .858, RMSEA = .066, SRMR = .056).

The item contents and the factor loadings of the four-factor CFA are presented in Table 6. In addition, the inter-correlation matrix of the four-factor CFA is displayed in Table 7.

Table 6
Item Factor Loadings for the Four-factor Model: Confirmatory Factor Analysis

Factor and item	Factor 1	2	3	4
Factor 1: Anxiety-Somatization				
10. I feel fearful.	.752			
9. I feel weak.	.723			
36. I feel nervous.	.682			
33. I feel that something bad is going to happen.	.663			
29. My heart pounds too much.	.550			
41. I have trouble falling asleep or staying asleep.	.522			
27. I have an upset stomach.	.493			
45. I have headaches.	.397			
Factor 2: Negative Self-worth				
31. I am satisfied with my life.		.824		
13. I am a happy person.		.786		
24. I like myself.		.745		
43. I am satisfied with my relationships with others.		.727		
20. I feel loved and wanted.		.720		
18. I feel lonely.		.679		
15. I feel worthless.		.674		
21. I enjoy my spare time.		.636		
37. I feel my love relationships are full and complete.		.549		
1. I get along with others.		.464		
Factor 3: Substance Abuse				
26. I feel annoyed by people who criticize my drinking or drug use.			.746	
32. I have trouble at work/school because of drinking or drug use.			.612	
11. After heavy drinking, I need a drink the next morning to get going.			.609	
Factor 4: Loss of Interest				
3. I feel no interest in things.				.717
28. I am working/studying less well than I used to.				.617
12. I find my work/school satisfying.				.605
22. I have difficulty concentrating.				.599
38. I feel that I am not doing well at work/school.				.588

Table 7
Factor Inter-correlation Matrix: Confirmatory Factor Analysis of the Four-factor Model

Factor	1	2	3
2	.651		
3	.141	.129	
4	.674	.675	.251

FACTOR MIXTURE MODELING

A total of 12 alternative factor mixture models with increasing numbers of factors and classes were fitted. The three covariates of sex, clinical status, and race were not added to these models. The 12 models can be grouped into three categories based on the number of factors: one-, three-, and four-factor CFA models. Also, each category has four models with increasing number of classes from one to four. After the models were fitted to the second subsample of the data (N = 1,413), the fit indices of these alternative models were compared to decide which model(s) fit best to the data within each model category (see the three upper panes of Table 8). In addition, the fourth pane of Table 8 shows the fit indices of a revised three-factor model that was resulted after three items representing the Substance Abuse factor was deleted from the four-factor model. Thus, the revised three-factor model has only 23 items, while the one- and three-factor model have 45 items and the four-factor model has 26 items. The model fit indices of the revised three-factor model will be discussed later.

Table 8
Fit indices for 16 Factor Mixture Models without Covariates

Models	AIC	BIC	aBIC	aLRT (*p*-value)	Entropy	Loglikelihood
1f 1c	174,491.20	175,200.42	174,771.57	N/A	N/A	-87,110.60
2c	174,451.32	175,171.05	174,753.85	**.006**	.528	-87,088.66
3c	174,438.80	175,169.03	174,727.48	**.045**	.613	-87,080.40
4c	**174,427.08**	**175,167.82**	**174,719.91**	**.021**	**.682**	-87,072.54
3f 1c	163,426.46	164,104.16	163,694.37	N/A	N/A	-81,584.23
2c	163,366.30	164,065.02	163,642.52	**.005**	.588	-81,550.15
3c	163,326.30	**164,046.02**	163,610.82	.136	.698	-81,526.15
4c	**163,308.83**	164,049.57	**163,601.67**	.243	**.712**	-81,513.42
4f 1c	97,659.26	98,100.56	97,833.72	N/A	N/A	-48,745.63
2c	96,821.73	97,289.29	97,006.57	.096	**.969**	-48,321.87
3c	96,784.47	97,278.29	96,979.69	.505	.760	-48,298.23
4c	**96,090.63**	**96,610.72**	**96,296.24**	.290	.851	-47,946.32
3f[a] 1c	90,357.88	90,736.13	90,507.41	N/A	N/A	-45,106.94
2c	90,279.05	90,678.31	90,436.89	**.000**	.636	-45,063.53
3c	90,230.68	90,650.95	90,396.82	**.001**	**.766**	-45,035.34
4c	**90,193.84**	**90,635.14**	**96,368.30**	**.015**	.694	-45,012.92

Note. $N = 1{,}413$. 'f' represents factor(s) and 'c' class(es). The best index values under each category of factor mixture models and significant aLRT *p*-values are highlighted.
[a] This three-factor model resulted after the Substance Abuse factor was excluded from the four-factor model.

In model comparison, a model with smaller values on each of the AIC, BIC, and aBIC is preferred over a model with a larger value. Although loglikelihood values, presented in the last column, are used for model comparison and higher values are preferred, they are not used alone because they simply increase as more parameters (e.g., the number of classes) are added into a model. However, the information criteria such as AIC, BIC ,and aBIC, although based on loglikelihood, penalize the loglikelihood for the number of parameters estimated (McLachlan & Peel, 2000; Muthén, 2006). The values of

the entropy are presented in the fifth column. The entropy, which ranges 0 to 1 with 1 being optimal, functions as an indicator of how well separated estimated classes are from each other (Muthén, 2006). In addition, aLRT *p*-values are presented, providing a test where a statistically significant *p*-value supports the fit of the associated model with *k* classes over the fit of the same model but with *k-1* classes (Lubke & Muthén, 2005).

Among the one-factor factor mixture models (see the first pane of Table 8), the four-class model is found to fit better than the models with fewer classes as indicated by the values of AIC, BIC, and aBIC. Also, the *p*-values of the aLRT indicate that the four-class solution provides a better fit over the corresponding one-, two-, and three-class solutions. Unlike for the one-factor factor mixture models, the fit indices of three-factor FMMs produced more inconsistent results (see the second pane of Table 8). For example, the AIC and aBIC indicate that the four-class solution is the best, but the BIC supports the three-class model. Furthermore, the aLRT *p*-values support the better fit of the two-class solution. In the four-factor FMMs, the four-class solution exhibits the lowest values on the AIC, BIC, and aBIC. Nevertheless, the aLRT *p*-values support the single-class model over the other multi-class models.

All in all, considering all the twelve factor mixture models together, the four-factor FMMs have the best fit based on the values of the AIC, BIC, and aBIC, which are much smaller than those for either the one-factor or three-factor FMMs. Therefore, the four-factor FMMs were investigated further to identify the best-fitting model amongst them. It was found that the factor loadings on the three items of the Substance Abuse factor in these four-factor models were mostly very low. Specifically, the standardized factor loadings of the three items were .085, .094, and .194 in the two-class model; .097,

.113, and .222 in the three-class model; and .146, .995, and .145 in the four-class model. Therefore, while the content of the OQ items seemed to indicate they were measuring Substance Abuse, the indicators were not supported as good measures of the Substance Abuse factor. Thus only a revision of the four-factor FMM was explored in which these items were deleted resulting in a model using the 23 items that measured Anxiety-Somatization, Negative Self-worth, and Loss of Interest.

The revised three-factor FMMs with one to four classes were fitted to the data. All of the fit indices except the entropy supported the fit of the four-class solution (see the last pane of Table 8). The entropy, which indicates how distinct the classes are, supports the three-class solution. However, the information criteria (e.g., AIC, BIC, and aBIC), aLRT, and the entropy should not be the sole indicators of a model's fit. The composition of each class should be examined to help decide upon the best number of classes fitting a dataset (Lubke & Muthén, 2005). For example, although all fit indices might support a four-class over a three–class model, the fourth class might contain too few members to be interpreted as a meaningful class. Or, a well-interpretable, solid class might break into two uninterpretable classes as one more class is added to a factor mixture model. Furthermore, addition of a class should be supported by a relevant theory. For example, comparing classes in terms of factor means and covariates should make sense (Lubke & Muthén, 2005). Finally, it is important that the fit indices should not be the sole indicators determining model selection. As with EFA, where interpretation of factors is needed for model selection, so with FMM, interpretation of classes is essential for model selection (Allua, Stapleton, & Beretvas, 2007). For example, if classes cannot be interpreted or understood in FMM, then the classes are not useful.

Table 9 shows the class counts and proportions of the 16 factor mixture models that were compared. These numbers were calculated based on each subject's most likely membership. For example, the probabilities of belonging to each class in the three-factor, four-class model for the first subject in the dataset were .002, .171, .000, and .827 for the first through fourth classes, respectively. Therefore, the first subject was assigned to the most likely (fourth) class. However, the corresponding probabilities of the second subject in the same model were .089, .868, .000, and .043. Thus, the second subject was assigned to the second class.

Table 9
Class Counts and Proportions (FMM without Covariates)

Models	Class 1	Class 2	Class 3	Class 4
1f 2c	525 (.37)	888 (.63)		
3c	652 (.46)	143 (.10)	618 (.44)	
4c	407 (.29)	92 (.07)	521 (.37)	393 (.29)
3f 2c	467 (.33)	946 (.67)		
3c	42 (.03)	854 (.60)	517 (.37)	
4c	148 (.10)	32 (.02)	566 (.40)	667 (.47)
4f 2c	147 (.10)	1,266 (.90)		
3c	239 (.17)	143 (.10)	1,031 (.73)	
4c	142 (.10)	147 (.10)	955 (.68)	169 (.12)
3f[a] 2c	385 (.27)	1,028 (.73)		
3c	1,044 (.74)	351 (.25)	18 (.01)	
4c	221 (.16)	523 (.37)	19 (.01)	650 (.46)

Note. $N = 1{,}413$. 'f' represents factor(s) and 'c' class(es). Class counts and proportions are based on individuals' most likely latent class memebership. [a] This three-factor model resulted after the Substance Abuse factor was excluded from the four-factor model.

Further examination of the composition of Class 3 in the three- and four-class, revised three-factor FMMs revealed the numbers of Class 3 members was too small (i.e., n = 18 or 19) representing about 1% of the sample (see the last pane in Table 9). In order to investigate the stability of a given class transition matrices are utilized in which the membership of a *k-1*-class FMM is compared with that of a *k*-class model. Table 10 demonstrates that membership in classes in the three-class model can be traced back to that of the two-class model. For example, Class 3 of the three-class model is actually a small fragment from Class 2 of the two-class model. It can be further shown that Class 1 and Class 2 of the three-class model are almost identical with Class 2 and Class 1 of the two-class model, respectively. Therefore, the utility of the additional class in the three-class model is questionable. Table 11 shows the transition matrix of the three-class model into the four-class model in the revised three-factor FMM. It is also demonstrated that Class 2 (n =19) of the four-class model is almost identical to Class 3 (n = 17) of the three-class model. The other three classes of the four-class solution are different combinations of Classes 1 and 2 of the three-class solution.

Table 10
Change of Class Counts and Proportions: From Two Classes to Three Classes in the Revised Three-Factor FMM

Class	Class 1	Class 2	Class 3	Sum
1	37 (.03)	348 (.25)	0 (.00)	385 (.27)
2	1,007 (.71)	3 (.00)	18 (.01)	1,028 (.73)
Sum	1,044 (.74)	351 (.25)	18 (.01)	1,413 (1.00)

Note. N = 1,413. Class counts and proportions are based on individuals' most likely latent class membership. Columns correspond to the three classes fitted in the revised three-factor FMM without covariates, and rows correspond to the two classes fitted in the same FMM. The Sum row and column contain the marginal class counts of rows and columns.

Table 11
Change of Class Counts and Proportions: From Three Classes to Four Classes in the Revised Three-Factor FMM

Class	Class 1	Class 2	Class 3	Class 4	Sum
1	0 (.00)	2 (.00)	650 (.46)	392 (.28)	1,044 (.74)
2	221 (.16)	0 (.00)	0 (.00)	130 (.09)	351 (.25)
3	0 (.00)	17 (.01)	0 (.00)	1 (.00)	18 (.01)
Sum	221 (.16)	19 (.01)	650 (.46)	523 (.37)	1,413 (1.00)

Note. $N = 1{,}413$. Class counts and proportions are based on individuals' most likely latent class membership. Columns correspond to the three classes fitted in the revised three-factor FMM without covariates, and rows correspond to the two classes fitted in the same FMM. The Sum row and column contain the marginal class counts of rows and columns.

It seems that the two-class solution makes more sense than the three- and four-class solutions judging from the transition matrices that show one class in each solution (i.e., Class 3 in the three-class solution and Class 2 in the four-class solution) is very small in number. However, it should be noted transition matrices show only one thing. Using substantive theory to interpret class membership is also required to make a decision on the proper number of classes. Furthermore, although Class 3 in the three-class solution and Class 2 in the four-class solution are small in size, it is obvious the membership of these classes is almost identical, demonstrating the stableness of this class across the three- and four-class solutions. Thus, consulting relevant theory is necessary at this point.

Factor means and characteristics of each class in the two-, three-, and four-class models were investigated to see which factor mixture model is supported by the theory underlying the OQ-45. Table 12 shows the composition of each class with regard to sex, clinical status, and race. Table 12 also lists factor means for each class in each model. It

is hypothesized that factor means of the three factors will each point in the same direction so that they will all have either positive signs or all negative signs given that all the three factors for a single class. Remember that the three factors include: Anxiety-Somatization, Negative Self-worth, and Loss of Interest represent psychological distress or vulnerability. In other words, it would be unusual that one has anxiety and somatization, and happy about oneself concurrently. The factor means of the two-class model clearly demonstrates this hypothesized pattern. Note that the factor means of Class 2 (i.e., the reference class) of the two-class model are fixed to zero for model identification. All the factor means of Class 1 are significantly higher than those of Class 2 as hypothesized. However, as opposed to the hypothesis, the factor means of the three- and four-class models do not follow this pattern. For example, the factor means of Anxiety-Somatizatoion and Loss of Interest of Class 3 in the three-class model are the smallest, denoting the class is the most psychologically healthy one. Whereas, the factor mean of Negative Self-worth for this class was the highest, meaning the members of this class usually feels more unhappy and miserable than the members of the other two classes. This contradictory pattern of factor mean distribution is repeated among the factor means of the four-class model (see the last pane of Table 12). For example, the pattern of the factor means of Class 2 (n = 19) in this four-class model, which is identical to Class 3 (n = 17) of the three-class model indicate the members of Class 2 in the four-classs model shows least anxiety and somatization and function socially best, but they feel the most miserable and unhappy, which is very counter-intuitive.

Table 12
Class Proportions and Factor Means of Two-, Three-, and Four-class Models (Unconditional Model)

Class	Proportion Female	Clinical	White	Factor means (Standard errors) AS	NS	LI
				Total sample, N = 1,413		
	.43	.68	.81	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
				Two-class model		
1	.35	.89	.78	2.91* (0.16)	1.06* (0.13)	1.08* (0.14)
2	.46	.60	.82	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
				Three-class model		
1	.47	.61	.81	1.21* (0.39)	-1.93* (0.21)	1.04* (0.24)
2	.34	.89	.79	4.11* (0.43)	-0.93* (0.19)	2.12* (0.27)
3	.28	.56	.78	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
				Four-class model		
1	.30	.91	.81	3.40* (0.56)	-0.51† (0.21)	0.80* (0.18)
2	.26	.53	.84	-3.91* (0.74)	1.29* (0.27)	-1.53* (0.26)
3	.49	.50	.82	-2.98* (0.45)	-1.39* (0.00)	-0.75* (0.12)
4	.42	.81	.80	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)

Note. † $p < .05$. * $p < .01$. AS = Anxiety-Somatization; NS = Negative Self-worth; LI = Loss of Interest.

All in all, the transition matrices and the factor mean distributions indicate the two-class model fits the data best. Class 1 in the factor mixture model can be named the *Psychologically Vulnerable* class while Class 2 the *Psychologically Invulnerable* class. The composition of the two classes also can be analyzed based on the members' sex, clinical status, and race, which provides interesting findings. For example, the Psychologically Vulnerable class (i.e., Class 1) contains more males and more clinical subjects than the Psychologically Invulnerable class (i.e., Class 2). However, their racial composition was similar to each other (see Table 12). This indicates that the covariates of

sex and clinical status are a source of possible heterogeneity in the data. This suggests that these covariates be treated as known sources of heterogeneity and, subsequently, incorporated into the factor mixture model (Lubke & Muthén, 2005). Table 13 also indicates that the two covariates cause heterogeneity in the data. The subscale means of the three subscales (i.e., factors) are larger for women, and clinical population.

Table 13
Mean Subscale Scores for Sex, Clinical Status, and Race

Subscale	Female	Male	Clinical	Non-clinical	Non-White	White
AS (6.6)	10.4 (6.0)	12.0 (7.1)	12.8 (6.6)	8.2 (5.7)	11.6 (7.1)	11.2
NS (8.3)	14.4 (8.6)	16.6 (7.9)	18.2 (7.6)	10.4 (7.0)	16.3 (7.7)	15.5
LI (4.3)	9.3 (4.3)	9.9 (4.4)	10.5 (4.3)	7.8 (3.7)	10.5 (4.6)	9.4

Note. AS = Anxiety-Somatization; NS = Negative Self-worth; LI = Loss of Interest.

Next, a group of conditional factor mixture models where the three covariates of sex, clinical status, and race were included were fitted to the data. In this model, factor scores and the latent class variable were regressed on the covariates (see Figure 14 as an example) and their regression coefficients were estimated. In other words, these factor mixture models were fitted conditional on the three covariates. The revised three-factor model that consists of Anxiety-Somatization, Negative Self-worth, and Loss of Interest was incorporated as the within-class CFA model and four factor mixture models with one- to four-classes were fitted. Because these factor mixture models are conditional on the covariates, it is expected that the designated membership of two-, three-, and four-

class models is different from that of the unconditional models contained in Table 9 (Lubke & Muthén, 2005).

Table 14 displays fit indices of the four conditional factor mixture models. The information criteria and the aLRT *p*-value presented mixed findings. For example, the AIC and aBIC support the four-class solution, whereas the BIC and aLRT supported the two-class solution. Although the values of the entropy are similar across two-, three-, and four-class solutions, the three-class model has the largest value, indicating that the constituent classes of the three-class model are better separated from each other than those in the two- and four-class models. Transition matrices as well as factor intercepts and the covariate's regression weights on factor score were investigated to aid the process of selecting the model with the appropriate number of classes.

Table 14
Fit indices for Factor Mixture Models (Conditional Model with Covariates Included)

Models	AIC	BIC	aBIC	aLRT (*p*-value)	Entropy	Loglikelihood
3f 1c	89,979.53	90,405.06	90,147.76	N/A	N/A	-44,908.77
2c	89,845.74	**90,355.33**	90,047.19	**.001**	.680	-44,825.87
3c	89,762.05	90,355.69	89,996.73	.129	**.778**	-44,768.03
4c	**89,728.51**	90,406.21	**89,996.42**	.412	.663	-44,735.25

Note. $N = 1{,}413$. 'f' represents factor(s) and 'c' class(es). The best index values and a significant aLRT *p*-value are highlighted.

Tables 15 and 16 represent transition matrices for the two-class to three-class and three-class to four-class solutions. Overall, the matrices show that the classes in this conditional factor mixture model are less stable than those of the unconditional models (see Tables 10 and 11). However, some migration patterns can be observed. For example, most

members assigned to Class 2 in the three-class solution were drawn from Class 1 of the two-class solution. By the same token, about 86% (n = 605) of the total members of Class 3 in the three-class solution were drawn from Class 2 of the two-class solution. Also, most members of Class 3 of the four-class solution were drawn from Class 3 of the three-class solution.

Table 15
Change of Class Counts and Proportions: From Two Classes to Three Classes (Factor Mixture Models with Covariates)

Class	Class 1	Class 2	Class 3	Sum
1	215 (.15)	429 (.30)	66 (.05)	710 (.50)
2	77 (.05)	21 (.02)	605 (.43)	703 (.50)
Sum	292 (.21)	450 (.32)	671 (.48)	1,413 (1.00)

Note. N = 1,413. Class counts and proportions are based on individuals' most likely latent class membership. Columns respond to the three classes fitted in the three-class factor mixture model with covariates, and rows respond to the two classes fitted in the two-class factor mixture model. Sum represents marginal counts of rows and columns.

Table 16
Change of Class Counts and Proportions: From Three Classes to Four Classes

Class	Class 1	Class 2	Class 3	Class 4	Sum
1	36 (.03)	248 (.18)	7 (.01)	1 (.00)	292 (.21)
2	5 (.00)	25 (.02)	1 (.00)	419 (.30)	450 (.32)
3	31 (.07)	98 (.07)	366 (.26)	114 (.08)	671 (.48)
Sum	134 (.10)	371 (.26)	374 (.27)	534 (.38)	1,413 (1.00)

Note. N = 1,413. Class counts and proportions are based on individuals' most likely latent class memebership. Columns respond to the four classes fitted in the conditional factor mixture model and rows respond to the three classes fitted in the same factor mixture model. Sum represents marginal counts of rows and columns.

Next, factor intercepts and the regression coefficients of covariates on factor scores in each class were investigated (see Tables 17 and 18). Note that factor means

cannot be directly compared in a conditional model, where factor scores are regressed on covariates and, as a result, the factor means come to contain residual factor scores (see Lubke & Muthén, 2005 for more technical details; see also Equations 9 and 10). In lieu of factor means, factor intercepts and the regression coefficients of covariates on factor scores, the other two components in Equation 10 that are error-free, are compared together across classes. For example, in the two-class solution, the factor intercepts on Anxiety-Somatization and Negative Self-worth are higher in Class 1 than in Class 2, indicating the members of Class 1 are expressing more psychological distress. Also, the regression weights of the three covariates on the three factor scores were explored. For example, the standardized regression weight of clinical status on the Anxiety-Somatization factor for the non-clinical sample is -.896 and statistically significant. For the clinical sample the weight is 0 because the clinical sample was coded as 0 and the non-clinical sample 1. This shows that in Class 1 a non-clinical member typically scores .896 points lower on Anxiety-Somatization than a clinical member of the same sex and race. This prediction is in agreement with the expectation that a clinical person's level will be higher on the Anxiety-Somatization factor than a non-clinical person's one. This pattern is repeated with the factor score on the Negative Self-worth and Loss of Interest factors. However, this tendency is reversed for Class 2 at least with the Anxiety-Somatization factor scores, where a non-clinical member typically gets .378 points more than a clinical member of the same sex and race. Surprisingly, this finding is against a relevant theory that a clinical member scores higher on the factors of anxiety and somatization. All in all, as in the unconditional model, Class 1 can be termed the *Psychologically Vulnerable* class and Class 2 the *Psychologically Invulnerable* class.

Table 17
Class Proportions and Factor Intercepts of Two-, Three-, and Four-class Models (FMMs with Covariates)

Class	Proportion Female	Clinical	White	Factor intercepts (Standard error) AS	NS	LI
				Total sample, N = 1,413		
	.43	.68	.81	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
				Two-class model		
1	.42	.39	.83	2.75* (0.31)	0.78* (0.25)	0.22 (0.30)
2	.44	.97	.79	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
				Three-class model		
1	.38	.95	.86	2.57* (0.37)	0.54 (0.51)	1.01* (0.31)
2	.60	.04	.86	0.33 (0.88)	1.70* (0.42)	-1.73* (0.38)
3	.34	.99	.75	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
				Four-class model		
1	.81	.92	.81	0.30 (0.76)	0.18 (0.98)	1.91† (0.78)
2	.35	.92	.75	1.90* (0.51)	-0.06 (0.61)	1.26 (0.78)
3	.29	.97	.77	-1.46* (0.57)	-2.36* (0.59)	0.21 (0.71)
4	.49	.24	.88	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)

Note. † $p < .05$. * $p < .01$. AS = Anxiety-Somatization; NS = Negative Self-worth; LI = Loss of Interest.

Table 18
Regression Weights of the Covariates on the Factor Score

	Clinical status	Sex	Race
		Two-class model	
Class 1			
AS	-.896*	-.131*	.011
NS	-.634*	.037	-.054
LI	-.605*	-.005	.019
Class 2			
AS	.378*	-.035	-.029
NS	-.100	-.083	.073
LI	-.011	-.129†	-.164*

(table continues)

Table 18 (continued)

	Clinical status	Sex	Race
	Three-class model		
Class 1			
AS	.079	.317	-.135
NS	-.276†	-.111	.142
LI	-.163	-.041	-.148
Class 2			
AS	-.154	-.299*	-.019
NS	-.516*	.063	-.093†
LI	.110	.044	.042
Class 3			
AS	-.231*	.246*	-.135
NS	.191*	.063	-.046
LI	-.203*	-.021	-.164†
	Four-class model		
Class 1			
AS	.408*	.825*	-.198
NS	-.562*	-.505	-.058
LI	-.239	-.195	-.403†
Class 2			
AS	-.270	.095	.204
NS	-.193	.016	.170
LI	-.251†	-.016	.101
Class 3			
AS	-.276*	.305*	.066
NS	.433*	.223	.010
LI	-.232*	.081	-.162
Class 4			
AS	-.525*	-.256*	.071
NS	-.735*	.086	-.056
LI	-.298	.081	.037

Note. † $p < .05$. * $p < .01$. AS = Anxiety-Somatization; NS = Negative Self-worth; LI = Loss of Interest.

In the three-class model, a close examination of factor intercepts revealed an inconsistent relationship between the factor scores of Negative Self-worth and Loss of Interest (Table 17). In Class 2, one's level on Negative Self-worth is expected to be

highest across the classes, but one's level on Loss of Interest lowest, contradicting the conceptualization that one's sense of negative self-worth is negatively related to one's level of interest in things such as work and school. This incompatibility does not change even if the regression weights of covariates on the scores of Negative Self-worth and Loss of Interest are taken into account. For example, in Class 2 in the three-class model, a non-clinical, white person, controlling for one's sex, would score 1.091 on the factor of Negative Self-worth on average, which is still expected to be the highest across classes in the three-class solution, while his or her level on Loss of Interest is expected to be the lowest across classes, supported by no significant regression weights (see the second pane in Table 17). The four-class model looks more complicated. In addition, the three factors do a poor job at differentiating the four classes as evidenced by the fact that only four of the nine estimated factor intercepts are statistically significant (see the third pane of Table 17). Furthermore, the three covariates function poorly in the four-class model. For example, in Class 2, only one regression coefficients (i.e., clinical status on Loss of Interest) out of nine is statistically significant.

Furthermore, the usefulness of including covariates in the conditional factor mixture models can be explored through investigating the predictive power of covariates in class membership. It was shown that covariates predict the probability that a given subject belongs to a certain class (Equation 5). It is hypothesized, for example, in the two-class conditional model, that the probability that a member of the clinical sample belongs to the psychologically vulnerable class (i.e., Class 1) would be higher than the probability that the same member belongs to the psychologically invulnerable class (i.e., Class 2). However, surprisingly, this hypothesis was not supported by the result. First,

only 39% of Class 1 are from the clinical sample, while 97% of Class 2 are from the clinical sample (see Table 17). If the hypothesis were supported by the data, the proportion would be reversed.

In addition, the usefulness of incorporating covariates into the factor mixture model can be investigated with logistic regression coefficients. Table 19 shows estimated values of the logistic regression coefficients of Equation 5, where the categorical latent variable is regressed on the three covariates. In the first pane of Table 19, which shows the coefficients for the two-class model, the logistic regression coefficient for the clinical status covariate was 4.269 ($p < .05$). This coefficient denotes the log odds of the probability of belonging to Class 1 compared with the probability of belonging to the reference class (here, Class 2). Concretely, the log odds value of 4.269 suggests that the odds ratio for being in Class 1, the psychologically vulnerable class, is 73.591 (i.e., exp 4.269) times higher for the non-clinical sample than the clinical sample, controlling for the other two covariates (see Muthén, L. K., & Muthén, B. O., 1998-2007, for a more thorough explanation). This result is a direct opposite of the hypothesis that the probability of belonging to Class 1 for the clinical sample must be higher than the same probability for the non-clinical sample.

Table 19
Estimates of Logistic Regression Coefficients of the Categorical Latent Variable on the Covariates

	Estimate	*SE*	Estimate/*SE*	Odds ratio
		Two-class model		
Class 1				
Intercept	-1.871*	0.589	-3.178	N/A
Clinical Status	4.269*	0.809	5.275	71.470
Sex	1.419*	0.434	3.273	4.135
Race	0.032	0.309	0.105	1.033
		Three-class model		
Class 1				
Intercept	-1.174	0.745	-1.576	N/A
Clinical Status	1.524†	0.650	2.344	4.590
Sex	-0.159	0.412	-0.385	0.853
Race	0.657	0.528	1.244	1.929
Class 2				
Intercept	-4.195*	1.206	-3.479	N/A
Clinical Status	7.133*	1.043	6.836	1,252.700
Sex	-0.012	0.590	-0.020	0.988
Race	1.205†	0.537	2.244	3.338
		Four-class model		
Class 1				
Intercept	1.570	1.565	1.003	N/A
Clinical Status	-3.773*	1.427	-2.645	0.023
Sex	-2.192	1.129	-1.941	0.112
Race	-0.842	0.709	-1.187	0.431
Class 2				
Intercept	1.822	1.366	1.334	N/A
Clinical Status	-3.112†	1.218	-2.555	0.045
Sex	-0.424	0.583	-0.727	0.654
Race	-1.053	0.556	-1.894	0.349
Class 3				
Intercept	1.713	1.044	1.640	N/A
Clinical Status	-4.553*	1.061	-4.292	0.011
Sex	-0.283	0.763	-0.371	0.753
Race	-0.970†	0.463	-2.097	0.379

Note. † $p < .05$. * $p < .01$. In each model, the last class (e.g., Class 4 in the four-class model) represents the reference class.

The possible reasons of this reversed relationship between the latent class variable and the covariate of clinical status were investigated by comparing the observed scores on the three subscales among clinical and non-clinical groups in each class (Table 20). Table 20 demonstrates an interesting pattern. Class 2, the psychologically invulnerable class, draws 97% of its members from the clinical sample whose observed scores on the Anxiety-Somatization subscale are very low despite their clinical status. It indicates that although Class 2 appears to be dominated by the members from the clinical sample that are similar to their non-clinical counterparts in terms of their level on the Anxiety-Somatization subscale. This pattern is also repeated on the observed scores of the Loss of Interest subscale.

Table 20
Means and Standard Deviations of Observed Scores on the Three Subscales in Each Category of Class by Clinical Status

	AS	NS	LI
Class 1			
Clinical	20.2 (3.8)	21.5 (6.8)	12.3 (4.2)
Non-clinical	7.2 (4.6)	9.7 (6.5)	7.3 (3.3)
Class 2			
Clinical	9.5 (4.4)	16.7 (7.4)	9.7 (4.1)
Non-clinical	20.9 (3.4)	13.8 (9.4)	10.8 (4.0)

Note. $N = 1{,}413$. The values represent means followed by standard deviations in parenthesis.

In conclusion, in the factor mixture models estimated conditional on the three covariates, the two-class model prevails over the others as in the unconditional model. First, it is the most parsimonious model with the least number of parameters involved and is easy to comprehend. Second, its estimated parameters are simple and straightforward

to understand, which is not the case in the three- and four-class solutions. Third, not only the BIC but also the aLRT *p*-value support the two-class solution. BIC is known to impose a heavier penalty on complex models than AIC. Also, the BIC takes into consideration the sample size, which the AIC does not (McLachlan & Peel, 2000). However, it is questionable whether the three covariates should be included in the factor mixture model. First of all, the values of the information criteria and loglikelihood do not show much difference between the unconditional and conditional models (compare the last panes of Table 8 and Table 14). Second, when the values of the various regression coefficients (Tables 18 and 19) were investigated, many of them are not statistically significant or, if significant, the results are different than expected (see Table 20). Therefore, it is necessary to compare the unconditional, two-class FMM and the conditional, two-class FMM based on the revised three factor model further with a new sample and investigate whether the covariates provide any valuable information in understanding the heterogeneity of the data with that new data.

CROSS-VALIDATION

The two unconditional and conditional FMMs that are based on the revised three-factor model were cross-validated using Sample 3, the third subset of the three randomly divided subsamples. The purpose of the cross-validation was to see whether the general findings based on the original sample (i.e., Sample 2) were replicable with the new sample. Table 21 shows the fit indices for both the unconditional and conditional two-class FMMs using the cross-validation sample's scores. The differences in these values are very small and negligible. Also, the aLRT statistics suggest that both of the cross-

validated, two-class solutions fit better than their corresponding single-class models at the .05 level.

Table 21
Comparison of Fit Indices for FMMs Estimated with the Original and Cross-validated Samples

Models	AIC	BIC	aBIC	aLRT (*p*-value)	Entropy	Loglikelihood
Original sample						
UC	90,279.05	90,678.31	90,436.89	.000	.636	-45,063.53
C	89,845.74	90,355.33	90,047.19	.001	.680	-44,825.87
Cross-validation sample						
UC	90,506.10	90,905.36	90,663.94	.015	.581	-45,177.05
C	90,128.14	90,637.73	90,329.59	.033	.614	-44,967.07

Note. $N = 1{,}413$. All the models above are two-class, revised three-factor FMMs. UC = unconditional model. C = conditional model.

The number of members assigned to each class and their proportions in the cross-validated models were computed and compared with those from the models fitted on the original sample (Table 22). As expected, there were some changes in the membership across the samples although no dramatic change in class memberships was detected. For example, the membership of Class 1 in the unconditional model increased from 27% ($n = 385$) to 32% ($n = 446$), but the increment was relatively small. However, it should be noted that the process of classification of subjects into different classes in the estimation of a FMM is arbitrary. Therefore, Class 1 of the unconditional, two-class FMM fitted on the original sample (i.e., Sample 2) could correspond with Class 2 of the same model cross-validated using Sample 3. Thus, it is necessary to explore the membership of each class by comparing class membership proportional breakdowns, patterns of factor means

and using covariates in order to look into whether the two Class 1s are comparable to each other.

Table 22
Class Counts and Proportions

Models	Class 1	Class 2
Original sample		
UC	385 (.27)	1,028 (.73)
C	710 (.50)	703 (.50)
Cross-validation sample		
UC	446 (.32)	967 (.68)
C	593 (.42)	820 (.58)

Note. $N = 1{,}413$. Class counts and proportions are based on individuals' most likely latent class memebership. UC = unconditional model. C = conditional model.

Proportions of the subjects according to their clinical status, sex, and race in the cross-validated sample were calculated and compared with the proportions from the model fitted on the original sample (Table 23). Table 23 demonstrates that Class 1 of the unconditional model on the cross-validation sample is very similar to Class 1 of the FMM fitted on the original sample. For example, the proportions are almost identical across the two Class 1s. Also, the three factor means in the two classes are significantly different from zero and higher than those in Class 2s (see the middle two panes of Table 23). Therefore, Class 1 for the cross-validation sample corresponds closely with Class 1 on the original sample. Furthermore, Class 1 for the cross-validation sample can also be inferred as the psychologically vulnerable class and Class 2 the psychologically invulnerable class, just in the original sample.

Table 23
Class Proportions and Factor Means of the Two-class Conditional and Unconditional Models

Class	Proportion Female	Clinical	White	Factor intercepts[a] (Standard errors) AS	NS	LI
				Sample 2, N = 1,413		
	.43	.68	.81	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
				Sample 3 (Cross-validation sample), N = 1,413		
	.41	.68	.81	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
				Unconditional model: Original sample		
1	.35	.89	.78	2.91* (0.16)	1.06* (0.13)	1.08* (0.14)
2	.46	.60	.82	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
				Unconditional model: Cross-validation sample		
1	.33	.89	.80	2.67* (0.20)	1.02* (0.18)	0.97* (0.22)
2	.45	.58	.82	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
				Conditional model: Original sample		
1	.42	.39	.83	2.75* (0.31)	0.78* (0.25)	0.22 (0.30)
2	.44	.97	.79	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
				Conditional model: Cross-validation sample		
1	.35	.98	.81	-2.44* (0.35)	-0.45 (0.30)	-0.35 (0.48)
2	.46	.46	.81	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)

Note. † $p < .05$. * $p < .01$. AS = Anxiety-Somatization; NS = Negative Self-worth; LI = Loss of Interest. [a] The values of factor means are represented by those of factor intercepts in the two-class unconditional model.

For the conditional model estimated using the cross-validation sample, Class 1 looks very similar to Class 2 in the two-class, conditional model fitted using the original sample. Similarly, Class 2 of the cross-validation sample appears to share many characteristics of Class 1 of the original sample (see the last two panes of Table 23). For example, the class proportions of Class 1 in the original sample are similar to those of Class 2 of the cross-validated sample. Judging based on the values of the three factor

means it is also evident that the two classes are switched for the original sample and the cross-validation sample. For example, the signs are reversed between Class 1 of the original sample, representing a higher level of psychological vulnerability, and Class 1 of the cross-validation sample, representing a lower level of psychological vulnerability. Note that the factor means of the reference class (i.e., Class 2s in the last two panes) are fixed to zero due to model identification. Based on the factor intercepts as well as class proportions, Class 1 of the cross-validation sample can be called the psychologically invulnerable class and Class 2 the psychologically vulnerable class. Interestingly, as in the FMMs fitted with the original sample, Class 1 of the conditional model on the cross-validation sample (see the last pane of Table 23) draws most of its members from the clinical population, suggesting the covariate of clinical status does not explain heterogeneity in the data the way it is expected by substantive theory.

As stated before, in a conditional model, due to factor residual scores, both factor intercepts and regression weights of covariates on the factor scores should be considered together to examine the level of factors in each class. Table 24 shows the regression coefficients of the three covariates on the factor scores. Interestingly, the values of the regression coefficients on the original sample are very close to those on the cross-validation sample. The regression coefficient estimates for clinical status in Class 1 on the original sample (see the upper pane in Table 24) are almost the same as the corresponding estimates for the same covariate in Class 2 for the cross-validation sample (see the lower pane in Table 24). The estimates for the other covariates were also replicated closely across the original and cross-validation samples (see Table 24).

Table 24
Regression Weights of the Covariates on the Factor Score in the Two-class Conditional Model

	Clinical status	Sex	Race	Clinical status	Sex	Race
			Conditional model: Original sample			
Class 1				Class 2		
AS	-.896*	-.131	.011	.378*	-.035	-.029
NS	-.634*	.037	-.054	-.100	-.083	.073
LI	-.605*	-.005	.019	-.011	-.129†	-.164*
			Conditional model: Cross-validation sample			
Class 1				Class 2		
AS	.430*	.096	-.075	-.854*	-.074	.039
NS	.045	-.037	-.116†	-.643*	-.009	-.021
LI	.076	-.055	-.266*	-.586*	.007	.005

Note. † $p < .05$. * $p < .01$. AS = Anxiety-Somatization; NS = Negative Self-worth; LI = Loss of Interest.

Also, the logistic regression coefficients of the categorical latent variable on the three covariates were compared between the model fitted on the original sample and the same model estimated using the cross-validation sample (see Table 25). In Table 25, the coefficients of the psychologically vulnerable class (Class 1 for the original sample and Class 2 for the cross-validation sample) are compared. Although the values of the regression coefficients in the model were not exactly the same, a general pattern is still visible. The coefficients for the covariate of clinical status were still statistically significant for the cross-validated sample. However, while the sex variable's intercept and slope were statistically significant in the original sample, they were not with the cross-validation sample.

The coefficient for clinical status estimated with the cross-validation sample is interpreted in the same way as mentioned for the model fitted using the original sample. For example, the log odds of the probability of belonging to the psychologically vulnerable class compared with the probability of belonging to the psychologically invulnerable class is 3.460 times higher for the non-clinical sample than for the clinical sample. The log odds of 3.460 is equivalent to the odds of 36.817, indicating the probability of belonging to Class 2, the psychologically vulnerable class, compared with the probability of belonging to Class 1, the psychologically invulnerable class, is 36.817 times higher for the non-clinical sample than for the clinical sample, controlling for the sex and race. This finding is counter-intuitive because a member of the clinical sample is more likely to belong to the psychologically vulnerable class rather than to the invulnerable class.

Table 25
Estimates of Logistic Regression Coefficients of the Categorical Latent Variable on the Covariates (Two-class Conditional Model)

	Estimate	*SE*	Estimate/*SE*	Odds ratio
Conditional model: Original sample				
Class 1				
Intercept	-1.871*	0.589	-3.178	N/A
Clinical Status	4.269*	0.809	5.275	71.470
Sex	1.419*	0.434	3.273	4.135
Race	0.032	0.309	0.105	1.033
Conditional model: Cross-validation sample				
Class 2				
Intercept	-0.398	0.650	-0.612	N/A
Clinical Status	3.460*	1.012	3.420	36.817
Sex	0.384	0.286	1.342	1.468
Race	-0.314	0.453	-0.692	0.731

Note. * $p < .01$. In the upper pane, Class 2 is the reference class in which the coefficients are fixed to zero. However, in the lower pane, Class 1 is the reference class.

Although the regression coefficient of clinical status is interpreted differently than expected, the clinical status covariate apparently induces some of the heterogeneity in the data. It should be noted that this covariate is the only one whose regression coefficient is statistically significant in the conditional model estimated using the cross-validation sample. In addition, the same result was found with the original sample. That is, the psychologically invulnerable class of the cross-validation sample (i.e., Class 1) consists dominantly of the members of the clinical group. On the other hand, the psychologically vulnerable class (i.e., Class 2) is made up largely of those of the non-clinical group. For example, 98% of Class 1 in the cross-validated sample is drawn from the clinical group, while 46% of Class 2 from the same group. If the proportion of the clinical group in the cross-validation sample, 68%, is taken into account (see Table 23), the proportion (i.e.,

46%) of the clinical group in Class 2, the psychologically vulnerable class, becomes considerably smaller. This does not agree with the hypothesis mentioned earlier that a majority of the clinical group should comprise the psychologically vulnerable class. Table 26 demonstrates the reason why this disagreement happens.

Although Class 1, the psychologically invulnerable class, consists predominantly of clinical subjects, their mean scores on the three subscales are lower than the mean scores of the other clinical subjects in Class 2. Overall, the mean score distribution pattern indicates that the mean scores of the clinical group in Class 1 are similar to those of the non-clinical group in Class 2. By the same token, the mean scores of the non-clinical group in Class 1 are similar to those of the clinical group in Class 2. This phenomenon that already happened in the previous analysis using the original sample suggests that the covariate of clinical status provides little information because the FMM appears to differentiate the subjects based on their scores on the items. In other words, the clinical status variable is determined too simply as it is based on whether one has ever come to the counseling center or not and does not represent one's psychological functioning.

Table 26
Means and Standard Deviations of Observed Scores on the Three Subscales in Each Category of Class by Clinical Status: Comparison of the Original Sample and the Cross-Validation Sample

	AS	NS	LI
Model: Original sample			
Class 1			
Clinical	20.2 (3.8)	21.5 (6.8)	12.3 (4.2)
Non-clinical	7.2 (4.6)	9.7 (6.5)	7.3 (3.3)
Class 2			
Clinical	9.5 (4.4)	16.7 (7.4)	9.7 (4.1)
Non-clinical	20.9 (3.4)	13.8 (9.4)	10.8 (4.0)
Model: Cross-validation sample			
Class 1			
Clinical	8.6 (4.1)	16.3 (7.8)	9.1 (4.1)
Non-clinical	21.8 (3.0)	20.2 (4.8)	11.2 (4.0)
Class 2			
Clinical	19.3 (4.0)	21.4 (6.5)	12.5 (3.9)
Non-clinical	7.6 (4.7)	10.2 (6.7)	7.6 (3.3)

Note. N = 1,413. The values represent means followed by standard deviations in parenthesis. AS = Anxiety-Somatization; NS = Negative Self-worth; LI = Loss of Interest.

All in all, the findings from the cross-validation sample also suggest the unconditional two-class FMM is preferred over the conditional two-class one.

Chapter 5: Discussion

The foremost aim of this study was to investigate the factorial validity of the Outcome Questionnaire using confirmatory factor analysis and factor mixture modeling. The developers of OQ-45 suggested that OQ-45 consists of either one- or three-factors (Lambert et al., 1996; Lambert, Gregersen, & Burlingame, 2004). However, the results of the study supported neither the one- nor the three-factor solution. The values of all the fit indices did not meet the recommended criteria. The other two previous CFA studies where the factor structure of scores on OQ-45 was investigated also showed comparable results of a misfit of the one- and three-factor models on data (see Beretvas & Kearney, 2003; Mueller, Lambert, & Burlingame, 1998).

The CFA results in this study strongly indicate that the factor structure of OQ-45 is different from what its developers intended it to be. The result suggests that the aggregation of the OQ-45 individual item scores into either subscale scores or a total score cannot be justified. If one wants to aggregate the scores of different items, one should have evidence that these items measure the same construct, which can be substantiated by a confirmatory factor analysis. However, the CFA study demonstrated the OQ-45 items cannot be grouped together to measure either one- or three-dimensions. For example, the standardized factor loadings of 11 items in the one-factor model are less than .40, indicating that each of these items explain less than 16% of the variance of the single factor, psychological distress. Also, the standardized residual variances of the 45 items in the one-factor model ranged from .365 to .996 (M = 0.716, SD = 0.161), indicating that a considerable amount of item variances (i.e., 71.6% on average) is not

accounted for by the single factor. In other words, many items of OQ-45 either measures something else other than what is intended or are heavily affected by measurement error.

Furthermore, the CFA results for the three-factor model were not much better than those for the one-factor model. Although the number of standardized factor loadings less than .40 diminished to six, the distribution of standardized residual variances for the 42 items included in the three-factor model was similar to that of the 45 items in the one-factor model. The standardized residual variances of the 42 items in the three-factor model ranged from .360 to .994 ($M = 0.681$, $SD = 0.161$), demonstrating that a significant amount of item variances (i.e., 68.1% on average) is not explained by the three factors of Symptom Distress, Interpersonal Relations, and Social Role Performance. Also, this distribution of standardized residual variances shows that there is a noticeable difference among the extent of explained variances among the 42 items.

All of these results suggest that the OQ-45 items do not measure the three related constructs typically assumed by users of OQ-45, which make us question the validity, especially the factorial validity, of OQ-45. It was discussed earlier in Chapter Two that validity is related to the accuracy and properness of inferences that are made from individuals' responses on a test (Kane, 2006) and that construct validity is presently considered as the entirety of validity theory overarching all other types of validity such as content and criterion validity (Kane, 2001; Zumbo & Rupp, 2004). According to Messick (1998), these other types of validity cannot operate independently but can only provide complementary information to construct validity. Construct validity refers to the extent to which inferences made from individuals' responses on a test appropriately capture the theoretical construct the test is intended to measure (Zumbo & Rupp, 2004). Factorial

validity, also a main focus of this study, was first proposed by Guilford (1946, p. 428) before the introduction of construct validity in the 1950s: "The factorial validity of a test is given by its loadings in meaningful, common, reference factors. This is the kind of validity that is really meant when the question is asked 'Does this test measure what it is supposed to measure?'... The answer then should be in terms of factors and their loadings." The results of the present study demonstrate that the scores on the OQ-45 do not have sufficient factorial validity for a good psychological measure, a finding that is consistent with the results of the two other CFA studies (Beretvas & Kearney, 2003; Mueller, Lambert, & Burlingame, 1998).

The weak factorial validity of OQ-45 suggests the need to alter current practices with OQ-45. Above all, the results suggest that the OQ-45 items do not measure either the single or the three related constructs typically assumed by users of OQ-45, rather the items assess something else. Therefore, it is not recommended that psychologists aggregate the scores of the individual items to calculate either a total score or three subscale scores as proposed. This recommendation has some practical implications. First, the cutoff score of 63 or 64 on the OQ-45 total score may not function as an appropriate standard to distinguish a clinical population from a non-clinical population. Second, the validity of the reliability change index (RCI) of 14 points, which was derived from a formula by Jacobson and Truax (1991), is weakened as an indicator of significant clinical change. Note that a client who improves 14 or more points in either positive or negative direction on the OQ-45 is considered as having made clinically significant change (Hannan et al., 2005; Lambert, Gregersen, & Burlingame, 2004).

Also, if the factorial validity of OQ-45 is in question, its extensive usage in clinical and research settings should be reconsidered because any utilization of a psychological measure requires empirically sound factorial validity. OQ-45 has been widely employed as a tool for monitoring treatment efficacy, for making informed decisions about clinically significant change, and for establishing psychotherapy goal criteria (Lambert, Gregersen, & Burlingame, 2004). If a test does not measure the construct that is intended to measure, its proposed applications will generate dubious results. For example, the questionable factorial validity of OQ-45 undermines the rationale for its proposed use of calculating expected recovery rate of clients (Lambert, Hansen, & Finch, 2001) and of a feedback algorithm for clients' clinical progress (Lambert, Whipple, Bishop, et al., 2002). In addition, due to the controversial factorial validity of OQ-45, the possibility of false positives or false negatives is increased; therefore, any clinical decisions based on OQ-45 may be erroneous and result in denial of services to those who really need them. For example, in one study, the RCI of 14 points on OQ-45 was used to classify clients into four different categories of deteriorated, no change, improved, and recovered (Hansen, Lambert, & Forman, 2002). If this categorization is ensued by further decisions that may impact clients' future lives (e.g., academic accommodation, reimbursement in managed care, or court decisions), these crucial decisions needs to be well supported by a measure of established validity. Furthermore, the disputed validity of OQ-45 challenges the pertinence of its use as an evaluation tool for performance of counseling center staff including interns and practicum students (e.g., Okiishi, Lambert, Nielsen, & Ogles, 2003).

The misfit of the one- and the three-factor models to the data in this study also indicates that scores on the other related versions of the Outcome Questionnaire such as the LSQ, YLSQ, and Y-OQ-30.1 should be examined again for their psychometric properties, especially their factorial validity. Scores on these questionnaires have also not been factor analyzed (e.g., EFA or CFA). Any effort to examine and improve their factorial validity is urgent because these measures are also widely used. For example, the LSQ, a 30-item version of OQ-45 has been extensively used by the PacifiCare Behavioral Health (PBH), a managed behavioral health care organization serving millions of members (Brown, Burlingame, Lambert, Jones, & Vaccaro, 2001; PacifiCare Behavioral Health, Inc., 2005). However, to the best of my knowledge, no information on the factor structure of the scores on the LSQ is available. It is all the more urgent because PBH employs LSQ and the Youth Life Status Questionnaire (YLSQ) or Y-OQ-30.1 for its clinical outcome management program (Brown, Burlingame, Lambert, Jones, & Vaccaro, 2001; Wampold & Brown, 2005) and evaluates clinical performance of service providers based on these measures (Brown & Jones, 2005).

In the meantime, the four-factor model of 26 OQ-45 items that has been identified through exploratory and confirmatory factor analyses in this study seems to hold some promise as an alternative. However, the shortened, 26-item version of the OQ-45 needs more validation prior to being used. First, the CFA result suggests that although the four-factor model was much improved over that of the one- and three-factor models it was still not found to fit the data. This suggests that a new, more rigorous factor model that will provide a better fit to data be identified. Second, other psychometric properties of the new four-factor model should be researched. As discussed in the literature review, factorial

validity is only one of many forms of validity required for a valid psychological measure. For example, convergent/discriminant validity, predictive validity, and consequential validity of the new four-factor model should be further studied before it is adopted in research and clinics.

This study also demonstrated the usefulness and flexibility of factor mixture modeling, which explains heterogeneity in data using not only continuous latent variables, but also categorical latent variables (Lubke & Muthén, 2005). Thus, a FMM is more flexible than a CFA model (i.e., a single-class FMM) in explaining variation in data. This study demonstrated the FMM always fits better than single-class CFA models. A typical CFA analysis cannot model the kind of heterogeneity in data that was uncovered through factor mixture modeling in this study.

Furthermore, according to the information criteria and aLRT *p*-values, the two-class models fit better than the three- and four-class FMMs. However, in factor mixture modeling, it should be noted how constituent classes are interpreted should be considered in addition to results from various fit indices when deciding on an appropriate number of classes (Allua, Stapleton, & Beretvas, in press). Also, relevant theory should support the interpretation of classes in terms of their composition and factor mean pattern.

For the current data, membership in the two-class FMMs further supported the fit of the models. Examination of classes in FMMs with different number of classes suggests that there are two heterogeneous classes in the data, where the same three-factor model holds in each class but factor means differ across classes. Remember, as in exploratory factor analysis, classes are arbitrary and are named after their characteristics (Allua, Stapleton, & Beretvas, in press). Therefore, the class where the factor means are

significantly higher was named the psychologically vulnerable class and the other class with lower factor means the psychologically invulnerable class.

In the unconditional, two-class FMMs that were fitted on both the original sample (i.e., Sample 2) and the cross-validation sample (i.e., Sample 3), all the factor means on the three factors were significantly higher in the psychologically vulnerable class than in the invulnerable class (see Table 21). Nevertheless, this pattern became slightly different once the three covariates were incorporated into the models. In the conditional model that was fitted using the original sample, the factor means on the Anxiety-Somatization and Negative Self-worth factors were significantly higher in the psychologically vulnerable class, but the factor means on the Loss of Interest factor was not significantly different across the two classes. However, in the conditional model that was fitted using the cross-validation sample, only the mean on the Anxiety-Somatization factor was significantly greater in the psychologically vulnerable class. At this point, one may be curious about the reason behind this difference of factor mean pattern between the unconditional model and the conditional model. It should be noted that in a conditional model the FMM is estimated not only based on the latent class variable but also conditional on covariates included (Lubke & Muthén, 2005).

Three covariates of clinical status, sex, and race were selected as known sources of heterogeneity and incorporated into the FMMs in this study. In factor mixture modeling, the latent class variable is regarded as explaining unknown sources of heterogeneity, and observed variables that are included into the model as covariates are considered to take account of known sources of heterogeneity (Lubke & Muthén, 2005). The three covariates, especially the clinical status variable, were sometimes efficient in

explaining heterogeneity across classes (see Table 24). For example, the clinical status variable explained about 80%, 40%, and 36% of the variances in the factor scores of Anxiety-Somatization, Negative Self-worth, and Loss of Interest, respectively, in the psychologically vulnerable class. However, in the psychologically invulnerable class, the three covariates, including the clinical status variable, explained a much smaller amount of variance in the factor scores. Furthermore, only three out of nine coefficients in each class in both the original and cross-validation samples were statistically significant (see Table 24). These results make one question the justification of including covariates into the FMM.

In addition, the three covariates were not so useful in explaining the between-class variability represented in the logistic regression of the latent class variable on the three covariates. For example, the clinical status variable was the only variable of which logistic regression coefficient was statistically significant and yielded a high odds ratio (see Table 25). However, the clinical status variable did a poor job of differentiating subjects when the subjects were assigned to either the psychologically vulnerable class or the psychologically invulnerable class (see Table 26). Table 26 clearly demonstrates that the subjects' responses on the 23 items held more responsibility for their membership assignment than their status on the clinical status variable. All in all, addition of the three covariates did not aid much in interpretation of the classes. For the same reasons, the two-class, unconditional FMM is preferred over the two-class, conditional FMM. Further studies are necessary where more known sources of heterogeneity are investigated and these observed sources are incorporated into FMMs that will investigate the factor structure of scores on the Outcome Questionnaire.

Estimation and comparison of FMMs is unique and thus more explanation on these topics is necessary. At least five or six assorted starting values that were specified by me were used for each FMM in this study. Then, the resulting loglikelihood values were compared and the model with the highest value of loglikelihood was chosen as the final result (see Muthén, L. K. & Muthén, B. O., 1998-2007 for more technical details). Only factor means were modeled as varying across classes in all the FMMs. All the remaining parameters, including intercepts, factor loadings, and error variances of items and factor variances were fixed equal across classes. Initially, I intended to relax the equality constraints to evaluate the potential heterogeneity of these parameters across classes. However, strict factorial invariance was instead assumed in the estimation of all FMMs because relaxing those constraints led to convergence problems. These convergence problems are prevalent in mixture model estimation (Muthén, L. K. & Muthén, B. O., 1998-2007). However, it is hoped this convergence problem will be overcome, for example, through introducing a more rigorous CFA model and by recruiting more participants. Also, unlike in a CFA model where various fit indices can be interpreted when evaluating the fit of a single model, it is not possible to investigate the fit of an individual FMM. There are, as of yet, no absolute criteria with which the fit of an individual FMM can be examined. A number of FMMs are usually fitted to data with an increasing number of classes and the relative fit of the different models is compared using fit indices such as information criteria, the entropy, and the aLRT *p*-value.

Because this study investigated the factor structure of scores on the OQ-45 at one time point (i.e., cross-sectional), it would be important for future research to examine

how the factor structure may be influenced by change over time. This is of particular importance because scores on the OQ-45 are intended to measure psychological change across psychotherapy sessions. Thus measurement invariance should hold not only among different groups but also across two or more measurement points (Meade, Lautenschlager, & Hecht, 2005). Otherwise, OQ-45 scores at two or more different time points should not be compared directly. In other words, the OQ-45 factor structure is required to be equivalent over time for scores on the scale to be used to assess change. Therefore, a further study is necessitated where longitudinal measurement invariance is investigated on OQ-45 scores. Growth mixture modeling, a form of factor mixture modeling, can be used to explore the issue of longitudinal measurement invariance.

This study clearly shows advantages of the mixture modeling framework. Researchers tend to stop their analysis and cease to go farther when the data under investigation does not fit the intended model. For example, if a certain CFA model does not fit the data at hand, we can think of many possibilities. The CFA model could be blamed for its poor representation of a latent construct. Or, given the fact that a CFA models depends on a specific sample, a distortion in sampling could be blamed. Although these hypotheses are well worth a further investigation to better understand a poor model fit, it is still possible that heterogeneity in data (e.g., mixture) may have caused the mismatch between the model and the data. In this case, if the mixture property is not appropriately taken account of, the true source of model misfit cannot be dealt with. With regard to this possibility, this study demonstrates that a FMM is superior, in terms of explaining variation across items, to traditional CFA models that cannot delve into heterogeneity in data. It was noted earlier that in spite of its flexibility for modeling

complex distributions, mixture modeling, except growth mixture modeling, has not been utilized much in psychology research. Mixture models have mostly been used in the natural sciences such as biology, astronomy, and genetics, and in the social sciences such as marketing and economics (McLachlan & Peel, 2000). It is strongly encouraged that mixture modeling be used more to assess heterogeneity in data in the field of counseling psychology as well as psychology in general.

The current study is based on a nationwide sample of over 4,000 college students and aims to promote mental and behavioral health of college students through an examination of factorial validity of a psychological measure widely used in college counseling centers. Concerns about the mental health of college students have grown in recent years (Sharkin & Coulter, 2005). The mental health status of this specific group was tremendously dramatized and given nationwide attention by high profile cases of campus shootings, such as the shooting at Northern Illinois University in February, 2008 and the shooting at Virginia Tech in April, 2007. Not only these immensely alarming incidents but also scholarly research indicates that college students' mental health appears to be declining. For example, Benton, Robertson, Tseng, Newton, and Benton (2003) investigated 13,257 student-clients of a college counseling center over 13 years from the late 1980s to the early 2000s and reported that the percentage of clients with 14 different clinical problems increased over the years. These 14 problems included serious ones that require more attention and resources such as depression, anxiety, suicidal ideation, sexual assault, and personality disorders as well as the "normal college student problems" such as developmental, situational, and relationship issues. The U.S. government also has recently demonstrated its interest in the mental welfare of college

students. President George W. Bush signed into law the Garrett Lee Smith Memorial Act in October 2004 (APA, 2004, October), which contains provisions from the Campus Care and Counseling Act (e.g., direct services and employing mental and behavioral health professionals). The Garrett Lee Smith Memorial Act was introduced by Senator Gordon Smith in remembrance of his son who had committed suicide, and represents "an important first step in establishing critical and needed support for mental and behavioral health services to students on college campuses" (APA, 2004, September). These new developments with regard to the mental health of college students require a more solid psychological questionnaire with which to make more informed clinical decisions and to monitor effective psychotherapy outcomes (Lambert, Gregersen, & Burlingame, 2004). It is hoped that this study can provide useful information on how a questionnaire with a solid validity should operate.

A few limitations with regard to the sample in the present study need to be acknowledged. First, the findings of this study from confirmatory factor analysis and factor mixture modeling are sample-dependent. Consequently, it would be important to determine whether the CFA models and factor mixture models tested in the study apply across different samples. Second, although fairly large and collected across the nation, the entire sample consists of college students, which may restrict the generalizability of the findings. It will be crucial to examine if the results of the study are replicated among other age groups. Third, the current study is based on a naturalistic sample. Naturalistic studies tend to show several methodological shortcomings, such as lack of experimental control, nonrandom assignment of participants, and confounding variables (Westen, Novotny, & Thompson-Brenner, 2004). However, this type of naturalistic sample should

provide useful information on how psychological constructs are actually conceptualized in the field (Seligman, 1995).

The main purpose of this study was to investigate the factor structure of the OQ-45 using a more advanced statistical tool of factor mixture modeling. Although the OQ-45 has enjoyed a wide popularity among professionals in the mental health field since its conception in mid 1990's, its factorial validity that was assessed with CFA was in question. An inadequate factorial validity would do a great harm to a psychological measure by decreasing confidence in the accurate reflection of the latent construct. This study combined CFA, a more traditional approach to study a factor structure of a measure, and FMM, a more advanced method dealing with a mixture in factor structure, in order to investigate the factor structure of the OQ-45. The results reinstated the results from two previous studies (Beretvas & Kearney, 2003; Mueller, Lambert, & Burlingame, 1998) that the supposed factor structure (i.e., one- or three-factor) of the OQ-45 is spurious. Furthermore, the newly derived revised four-factor solution did not provide an optimal fit to the data. It was later shown that models fit better to the data once the heterogeneity in data is accounted for with FMM. The results of FMM indicated that the unconditional, two-class solution is preferred to the one-class solution (i.e., traditional CFA model) and the data is better conceptualized as consisting of the psychologically vulnerable class and the psychologically invulnerable class. The study also showed the process of identifying the best fitting model cannot depend on fit indices alone but must also rely on interpretation of models and consideration of the ultimate use of the results.

Appendix A

THE OUTCOME QUESTIONNAIRE (OQ-45)

Instructions: Looking back over the last week, including today, help us understand how you have been feeling. Read each item and mark the oval under the category that best describes your current situation. For this questionnaire, work is defined as employment, school, housework, volunteer work, etc.

1. I get along with others.
2. I tire quickly.
3. I feel no interest in things.
4. I feel stressed at work/school.
5. I blame myself for things.
6. I feel irritated.
7. I feel unhappy in my marriage/significant relationship.
8. I have thoughts of ending my life.
9. I feel weak.
10. I feel fearful.
11. After heavy drinking, I need a drink the next morning to get going. (If you do not drink mark never)
12. I find my work/school satisfying.
13. I am a happy person.
14. I work/study too much.
15. I feel worthless.

16. I am concerned about family troubles.
17. I have an unfulfilling sex life.
18. I feel lonely.
19. I have frequent arguments.
20. I feel loved and wanted.
21. I enjoy my spare time.
22. I have difficulty concentrating.
23. I feel hopeless about the future.
24. I like myself.
25. Disturbing thoughts come into my mind that I can't get rid of.
26. I feel annoyed by people who criticize my drinking (or drug use). (If not applicable mark never).
27. I have an upset stomach.
28. I am working/studying less well than I used to.
29. My heart pounds too much.
30. I have trouble getting along with friends and close acquaintances.
31. I am satisfied with my life.
32. I have trouble at work/school because of drinking or drug use (If not applicable mark never).
33. I feel that something bad is going to happen.
34. I have sore muscles.
35. I feel afraid of open spaces, or of driving, or being on buses, subways, etc.
36. I feel nervous.

37. I feel my love relationships are full and complete.

38. I feel that I am not doing well at work/school.

39. I have too many disagreements at work/school.

40. I feel something is wrong with my mind.

41. I have trouble falling asleep or staying asleep.

42. I feel blue.

43. I am satisfied with my relationship with others.

44. I feel angry enough at work/school to do something I might regret.

45. I have headaches.

References

Agosti, V., Nunes, E., & Ocepeck-Welikson, K. (1996). Patient factors related to early attrition from an outpatient cocaine research clinic. *American Journal of Drug and Alcohol Abuse, 22*, 29-39.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317-332.

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Prospect Heights, IL: Waveland Press.

Allua, S., Beretvas, S. N., & Stapleton, L. M. (in press). Testing latent mean differences between observed and unobserved groups using multilevel factor mixture models. *Educational and Psychological Measurement.*

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Psychological Association. (2004, September). *Congressional update: Garrett Lee Smith Memorial Act passes House and Senate*. Retrieved June 8, 2008, from http://www.apa.org/ppo/issues/eglsupdt904.html

American Psychological Association. (2004, October). *The President signs the Garrett Lee Smith Memorial Act*. Retrieved June 8, 2008, from http://www.apa.org/ppo/issues/eglsbillsign.html

Arminger, G., Stein, P., & Wittenberg, J. (1999). Mixtures of conditional mean- and covariance-structure models. *Psychometrika, 64*, 475-494.

Arrindell, W. A., Barelds, D. P. H., Janssen, I. C. M., Buwalda, F. M., & van der Ende, J. (2006). Invariance of SCL-90-R dimensions of symptom distress in patients with peri partum pelvic pain (PPPP) syndrome. *British Journal of Clinical Psychology, 45*, 377-391.

Barkham, M., Evans, C., Margison, F., McGrath, G., Mellor-Clark, J., Milne, D., & Connell, J. (1998). The rationale for developing and implementing core outcome batteries for routine use in service settings and psychotherapy outcome research. *Journal of Mental Health, 7*, 35-47.

Bartholomew, D. J. (2002). Old and new approaches to latent variable modeling. In G. A. Marcoulides & I. Moustake (Eds.) *Latent variable and latent structure models* (pp. 1-13.). Mahwah, NJ: Lawrence Erlbaum.

Battle, C., Imbers, S., Hoen-Soric, R., Stone, A., Nash, C., & Frank, J. (1968). Target complaints as criteria of improvement. *American Journal of Psychology, 20,* 184–192.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods, 9*, 3-29.

Beck, A. (1972). Measuring depression: The depression inventory. In T. A.Williams, M. M. Katz, & J. A. Shield (Eds.), *Recent advances in the psychobiology of the depressive illness*. Washington, DC: U.S. Government Printing Office

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the BDI-II.* San Antonio, TX: Psychological Corporation.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychology, 4*, 53-6.

Benazzi, F. (2000). Female vs. male outpatient depression: A 448-case study in private practice. *Progress in Neuro-psychopharmacology & Biological Psychiatry, 24*, 475-481.

Benton, S. A., Robertson, J. M., Tseng, W-.C., Newton, F. B., & Benton, S. L. (2003). Changes in counseling center client problems across 13 years. *Professional Psychology: Research and Practice, 34*, 66-72.

Beretvas, S. N., & Kearney, L. K. (2003). *A shortened form of the Outcome Questionnaire: A validation of scores across groups.* A research report of the Research Consortium of Counseling and Psychological Services for Higher Education. Austin, TX: University of Texas at Austin, Counseling and Mental Health Center.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*, 203-219.

Brown, G. S., Burlingame, G. M., Lambert, M. J., Jones, E., & Vaccaro, J. (2001). Pushing the quality envelope: A new outcomes management system. *Psychiatric Services, 52,* 925-934.

Brown, G. S., & Jones, E. R. (2005). Implementation of a feedback system in a managed care environment: What are patients teaching us? *Journal of Clinical Psychology, 61*, 18-198.

Brown, G. S., Lambert, M. J., Jones, E., & Minami, T. (2005). Identifying highly effective psychotherapists in a managed care environment. *American Journal of Managed Care, 11,* 513-520.

Burlingame, G. M., Wells, M. G., Lambert, M. J., & Cox, J. C. (2004). Youth Outcome Questionnaire (Y-OQ). In M. E. Maruish, (Ed.), *The use of psychological testing for treatment planning and outcome assessment: Vol. 2. Instruments for Children and Adolescents* (3rd ed., pp. 235-273). Mahwah, NJ: Lawrence Erlbaum.

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-466.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Celeux G., & Soromenho G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13*, 195-212.

Colder, C. R., Campbell, R. T., Ruel, E., Richardson, J. L., & Flay, B. R. (2002). A finite mixture model of growth trajectories of adolescent alcohol use: Predictors and consequences. *Journal of Consulting and Clinical Psychology, 70*, 976-985.

Colder, C. R., Mehta, P., Balanda, K., Campbell, R. T., Mayhew, K. P., Stanton, W. R., Pentz, M. A., & Flay, B. R. (2001). Identifying trajectories of adolescent smoking: An application of latent growth mixture modeling. *Health Psychology, 20*, 127-135.

Connell, J., Barkham, M., & Mellor-Clark, J. (2007). CORE-OM mental health norms of students attending university counselling services benchmarked against an age-matched primary care sample. *British Journal of Guidance and Counselling, 35*, 41-57.

Cramer, K. M. (1999). Psychological antecedents to help-seeking behavior: A reanalysis using path modeling structures. *Journal of Counseling Psychology, 46*, 381-387.

Derogatis, L. R., & Cleary, P. A. (1977). Confirmation of the dimensional structure of the SCL-90: A study in construct validation. *Journal of Clinical Psychology, 33*, 981-989.

Derogatis, L. R., Rickels, K., & Rock, A. F. (1976). The SCL-90 and the MMPI: A step in the validation of a new self-report scale. *British Journal of Psychology, 128*, 280-289.

Dunn, T. W., Burlingame, G. M., Wells, M. G., Walbridge, M., Smith, J., & Crum, M. J. (2005). Outcome assessment for children and adolescents: Psychometric validation of the Youth Outcome Questionnaire 30.1 (Y-OQ-30.1). *Clinical Psychology and Psychotherapy, 12,* 388-401.

Ellickson, P. L., Martino, S. C., & Collins, R. L. (2004). Marijuana use from adolescence to young adulthood: Multiple developmental trajectories and their associated outcomes. *Health Psychology, 23*, 299-307.

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J., & Audin, K. (2002). Towards a standardized brief outcome measure: psychometric properties and utility of the CORE-OM. *British Journal of Psychiatry, 180*, 51-60.

Floyd, F. J., & Widman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7,* 286–299.

Froyd, J. E., Lambert, M. J., & Froyd, J. D. (1996). A review of practices of psychotherapy outcome measurement. *Journal of Mental Health, 5*, 11-16.

Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling, 3*, 62-72.

Greenbaum, P. E., Del Boca, F. K., Darkes, J., Wang, C., & Goldman, M. S. (2005). Variation in the drinking trajectories of freshman college students. *Journal of Consulting and Clinical Psychology, 73,* 229-238.

Gregersen, A. T., Nebeker, R. S., Seely, K. L., & Lambert, M. J. (2004). Social validation of the Outcome Questionnaire-45: An assessment of Asian and Pacific Islander college students. *Journal of Multicultural Counseling and Development, 32*, 194-205.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*, 427-439.

Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Hamilton, M. (1950). The assessment of anxiety states by rating. *British Journal of Medical Psychology, 32,* 50–55.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry, 23,* 56–61.

Hancock, G. R. (1997). Structural equation modeling methods of hypothesis testing of latent variable means. *Measurement and Evaluation in Counseling and Development, 30*, 91-105.

Hancock, G. R. (2004). Experimental, quasi-experimental, and nonexperimental design and analysis with latent variables. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 317-334). Thousand Oaks, CA: Sage.

Hancock, G. R., Lawrence, F. R., & Nevitt, J. (2000). Type I error and power of latent mean methods MANOVA in factorially invariant and noninvariant latent variable systems. *Structural Equation Modeling, 7*, 534-556.

Hancock, G. R., Stapleton, L. M., & Arnold-Berkovits, I. (2006). *The tenuousness of invariance tests within multisample covariance and mean structure models.* Unpublished manuscript.

Hankin, B. L., Abramson, L. Y., Moffitt, T. E., Silva, P. A., McGee, R., & Angell, K. E. (1998). Development of depression from preadolescence to young adulthood: Emerging gender differences in a 10-year longitudinal study. *Journal of Abnormal Psychology, 107*, 128-140.

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., & Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology, 61*, 155-163.

Hansen, N. B., Lambert, M. J., & Forman, E. M. (2002). The psychotherapy dose-response effect and its implications for treatment delivery services. *Clinical Psychology: Science and Practice, 9,* 329-343.

Harford, T. C., Wechsler, H., & Muthén, B. O. (2003). Alcohol-related aggression and drinking at off-campus parties and bars: A national study of current drinkers in college. *Journal of Studies on Alcohol, 64*, 704-711.

Hatfield, D. R., & Ogles, B. M. (2004). The use of outcome measures by psychologists in clinical practice. *Professional Psychology: Research and Practice, 35*, 485-491.

Hessen, D. J., Dolan, C. V., & Wicherts, J. M. (2006). The multigroup common factor model with minimal uniqueness constraints and the power to detect uniform bias. *Applied Psychological Measurement, 30*, 233-246.

Horowitz, L. M., Rosenberg, S. E., Baer, B. A., Ureño, G., & Vilaseñor, V. S. (1988). Inventory of interpersonal problems: Psychometric properties and clinical applications. *Journal of Consulting and Clinical Psychology,*56, 885-892.

Hu, L. –T. & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 1*, 1-55.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12-19.

Jackson, K. M., & Sher, K. J. (2005). Similarities and differences of longitudinal phenotypes across alternate indices of alcohol involvement: A methodologic comparison of trajectory approaches. *Psychology of Addictive Behaviors, 19*, 339-351.

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science, 16*, 39-59.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409-426.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70*, 631–639.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319-342.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education/Praeger.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, CA: Sage.

Kendler, K. S., Thornton, L. M., & Prescott, C. A. (2001). Gender differences in the rates of exposure to stressful life events and sensitivity to their depressogenic effects. *American Journal of Psychiatry, 158*, 587-593.

Kessler, R. C., McGonagle, K. A., Swartz, M., Blazer, D. G., & Nelson, C. B. (1993). Sex and depression in the National Comorbidity Survey. I: Lifetime prevalence, chronicity, and recurrence. *Journal of Affective Disorders, 29,* 85-96.

Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C., & Yanchar, S. C. (1996). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology and Psychotherapy, 3,* 249–258.

Lambert, M. J., Gregersen, A. T., & Burlingame, G. M. (2004). The Outcome Questionnaire. In M. E. Maruish, (Ed.), *The use of psychological testing for treatment planning and outcome assessment: Vol. 3. Instruments for Adults* (3rd ed., pp. 191-234). Mahwah, NJ: Lawrence Erlbaum.

Lambert, M. J., Hannöver, W., Nisslmüller, K., Richard, M., & Kordy, H. (2002). Fragebogen zum ergebnis von psychotherpie: Zur reliabilitat und validitat der deutschen ubersetzung des Outcome Questionnaire 45.2 (OQ-45.2). *Questionnaire on the results of psychotherapy: Reliability and Validity of the German translation of the Outcome Questionnaire 45.2 (OQ45.2). Zeitschrift für Klinische Psychologie und Psychotherapie,31,* 40-47.

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69,* 159-172.

Lambert, M. J., & Hawkins, E. J. (2004). Measuring outcome in professional practice: Considerations in selecting and using brief outcome instruments. *Professional Psychology: Research and Practice, 35*, 492-499.

Lambert, M. J., Lunnen, K., Umphress, V., Hansen, N. B., & Burlingame, G.M. (1994). *Administration and scoring manual for the Outcome Questionnaire (OQ-45.1).* Salt Lake City: IHC Center for Behavioral Healthcare Efficacy.

Lambert, M. J., Ogles, B. M., & Masters, K. S. (1992). Choosing outcome assessment devices: An organizational and conceptual scheme. *Journal of Counseling & Development, 70*, 527-532.

Lambert, M. J., Smart, D. W., Campbell, M. P., Hawkins, E. J., Harmon, C. & Slade, K. L. (2006). Psychotherapy outcome, as measured by the OQ-45, in African American, Asian/Pacific Islander, Latino/a, and Native American Clients compared with matched Caucasian clients. *Journal of College Student Psychotherapy, 20*, 17-29.

Lambert, M. J., Whipple, J. L., Bishop, M. J., Vermeersch, D. A., Gray, G .V., & Finch, A. E. (2002). Comparison of empirically-derived and rationally-derived methods for identifying patients at risk for treatment failure. *Clinical Psychology and Psychotherapy, 9,* 149-164.

Lambert, M. J., Whipple, J. L., Hawkins, E. J., Vermeersch, D. A., Nielsen, S. L., & Smart, D. W. (2003). Is it time for clinicians to routinely track patient outcome? A meta-analysis. *Clinical Psychology: Science and Practice, 10,* 288-301.

Levitt, H. M., Stanley, C. M., Frankel, Z., & Raina, K. (2005). An evaluation of outcome measures used in humanistic psychotherapy research: Using thermometers to weigh oranges. *The Humanistic Psychologist, 33*, 113-130.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*, 53-76.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*, 767-778.

Lubke, G. H., & Dolan, C. V. (2003). Can unequal residual variances across groups mask differences in residual means in the common factor model? *Structural Equation Modeling, 10*, 175-192.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). Weak measurement invariance with respect to unmeasured variables: An implication of strict factorial invariance. *British Journal of Mathematical and Statistical Psychology, 56*, 231-248.

Lubke, G. H., & Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*, 21-39.

Lubke, G. H., & Muthén, B. O. (2005). Perfomance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling, 14*, 26-47.

Magidson, J., & Vermunt, J. K. (2002). Latent class models for clustering: A comparison with K-means. *Canadian Journal of Marketing Research, 20*, 37-44.

Magidson, J., & Vermunt, J. K. (2004). Latent class models. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 175-198). Thousand Oaks, CA: Sage.

McCullough, M. E., Enders, C. K., Brion, S. L., & Jain, A. R. (2005). The varieties of religious development in adulthood: A longitudinal investigation of religion and rational choice. *Journal of Personality and Social Psychology, 89*, 78-89.

McLachlan, G., & Peel, D. (2000). *Finite mixture models.* New York: Wiley.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543.

Meade, A. W., Lautenschlager, G. J., & Hecht, J. E. (2005). Establishing measurement equivalence and invariance in longitudinal data with item response theory. *International Journal of Testing, 5*, 279-300.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Pyschometrika*, *58*, 525-543.

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change: Decade of behavior* (pp. 203-240). Washington, DC: American Psychological Association.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45*, 35-44.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*, 479-515.

Moore, W. L. (1980). Levels of aggregation in conjoint analysis: An empirical comparison. *Journal of Marketing Research, 17*, 516-523.

Mueller, R. M., Lambert, M. J., & Burlingame, G. M. (1998). Construct validity of the Outcome Questionnaire: A confirmatory factor analysis. *Journal of Personality Assessment, 70,* 248-262.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557-585.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*, 338-354.

Muthén, B. O. (2002). Byeond SEM: General latent variable modeling. *Behaviormetrika, 29*, 81-117.

Muthén, B. O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 345-368). Newbury Park, CA: Sage.

Muthén, B. O. (2006). Should substance use disorders be considered as categorical or dimensional? *Addiction, 101*, 6-16.

Muthén, B. O., & Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. *Addictive Behaviors, 31*, 1050-1066.

Muthén, B. O., & Asparouhov, T. (in press). Should substance use disorders be considered as categorical or dimensional? *Addiction.*

Muthén, B. O., Asparouhov, T., & Rebollo, I. (2006). Advances in behavioral genetics modeling using Mplus: Applications of factor mixture modeling to twin data. *Twin Research and Human Genetics, 9,* 313-324.

Muthén, L. K., & Muthén, B. O. (1998-2006). *Mplus user's guide* (4th ed.). Los Angeles, CA: Muthén& Muthén.

Muthén, L. K., & Muthén, B. O. (2007). Mplus (Version 5.0) [Computer Software]. Los Angeles, CA: Muthén & Muthén.

Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55*, 463-469.

Nebeker, R. S., Lambert, M. J., & Huefner, J. C. (1995). Ethnic differences on the Outcome Questionnaire. *Psychological Reports, 77*, 875-879.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2006). *Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study.* Manuscript submitted for publication.

Okiishi, J., Lambert, M. J., Nielsen, S. L., & Ogles, B. M. (2003). Waiting for supershrink: An empirical analysis of therapist effects. *Clinical Psychology and Psychotherapy, 10*, 361-373.

Orlando, M., Tucker, J. S., Ellickson, P. L., & Klein, D. J. (2004). Developmental trajectories of cigarette smoking and their correlates from early adolescence to young adulthood. *Journal of Consulting and Clinical Psychology, 72*, 400-410.

PacifiCare Behavioral Health, Inc. (2005). *About PacifiCare Behavioral Health*. Retrieved September 3, 2006, from http://www.pbhi.com/About_Us/AboutUs.asp

Pearson, K. (1894). Contribution to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London (Ser. A), 185*, 71-110.

Pearson, K. (1895). Contribution to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London (Ser. A), 186*, 343-414.

Pinsoff, W. M., & Catherall, D. R. (1986). The integrative psychotherapy alliance: Family, couple and individual therapy scales. *Journal of Marital and Family Therapy, 12,* 137–151.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practices, 16*, 9-13.

Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling* (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum.

Reinecke, J. (2006). Longitudinal analysis of adolescents' deviant and delinquent behavior: Applications of latent class growth curves and growth mixture models. *Methodology, 2*, 100-112.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333-343.

Seligman, M. E. P. (1995). The effectiveness of psychotherapy: The Consumer Reports study. *American Psychologist, 50*, 965-974.

Sharkin, B. S., & Coulter, L. P. (2005). Empirically supporting the increasing severity of college counseling center client problems: Why is it so challenging? *Journal of College Counseling, 8*, 165-171.

Shostrom, E. L. & Knapp, R. R. (1966). The relationship of a measure of self actualization (POI) to a measure of pathology (MMPI) and to therapeutic growth. *American Journal of Psychotherapy, 20,* 193–202.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229-239.

Spanier, G. (1976). Measuring dyadic adjustment: New scales for assessing the quality of marriage and similar dyads. *Journal of Marriage and the Family, 38,* 15–28.

Spielberger, C., Gorsuch, A., & Lushene, R. (1970). *The state-trait anxiety inventory.* Palo Alto, CA: Consulting Psychologist Press.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, New Jersey: Lawrence Erlbaum.

Stoolmiller, M., Kim, H. K., & Capaldi, D. M. (2005). The course of depressive symptoms in men from early adolescence to young adulthood: Identifying latent trajectories and early predictors. *Journal of Abnormal Psychology, 114*, 331-345.

Tucker, J. S., Orlando, M., & Ellickson, P. L. (2003). Patterns and correlates of binge drinking trajectories from early adolescence to young adulthood. *Health Psychology, 22*, 79-87.

Umphress, V. J., Lambert, M. J., Smart, D. W., Barlow, S. H., & Clouse, G. (1997). Concurrent and construct validity of the Outcome Questionnaire. *Journal of Psychoeducational Assessment, 15*, 40-55.

Vermeersch, D. A., Lambert, M. J., & Burlingame, G. A. (2000). Outcome Questionnaire: Item sensitivity to change. *Journal of Personality Assessment, 74*, 242-261.

Vermeersch, D. A., Whipple, J. L., Lambert, M. J., Hawkins, E. J., Burchfield, C. M., & Okiishi, J. C. (2004). Outcome Questionnaire: Is it sensitive to changes in counseling center clients? *Journal of Counseling Psychology, 51*, 38-49.

Wackerly, D. D., Mendenhall, W., III, & Scheaffer, R. L. (2002). *Mathematical statistics with applications* (6th ed.). Pacific Grove, CA: Duxbury.

Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: An illustration with the MMPI. *Psychological Methods, 5*, 125-146.

Wampold, B. E., & Brown, G. S. (2005). Estimating variability in outcomes attributable to therapists: A naturalistic study of outcomes in managed care. *Journal of Consulting and Clinical Psychology, 73*, 914-923.

Weissman, M. M., & Bothwell, S. (1976). Assessment of social-adjustment by patient self-report. *Archives of General Psychiatry, 33,* 1111–1115.

Wells, M. G., Burlingame, G. M., Lambert, M. J., Hoag, M. J., & Hope, C. A. Conceptualization and measurement of patient change during psychotherapy: Development of the outcome questionnaire and youth outcome questionnaire. *Psychotherapy: Theory, Research, Practice, Training, 33*, 275-283.

Westen, D., Novotny, C. M., & Thompson-Brenner, H. (2004). The empirical status of empirically supported psychotherapies: Assumptions, findings, and reporting in controlled clinical trials. *Psychological Bulletin, 130*, 631-663.

Whipple, J. L., Lambert, M. J., Vermeersch, D. A., Smart, D. W., Nielsen, S. L., & Hawkins, E. J. (2003). Improving the effects of psychotherapy: The use of early identification of treatment failure and problem-solving strategies in routine practice. *Journal of Counseling Psychology, 50*, 59-68.

White, H. R., Bates, M. E., & Buyske, S. (2001). Adolescence-limited versus persistent delinquency: Extending Moffitt's hypothesis into adulthood. Journal of *Abnormal Psychology, 110*, 600-609.

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology, 89*, 696-716.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp.281-324). Washington, DC: American Psychological Association.

Yung, Y.-F. (1997). Finite mixtures in confirmatory factor-analysis models. *sychometrika, 62*, 297-330.

Zumbo, B. D., & Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 317-334). Thousand Oaks, CA: Sage.

Zung, W. W. (1965). A self-rating depression scale. *Archives of General Psychiatry,* 12, 63-70.

Zung, W. W. (1971). A rating instrument for anxiety disorders. *Psychosomatics,* 6, 371-379.

Vita

Seong-Hyeon (Sung) Kim was born on March 21, 1970, in Suncheon, Cheonnam, Korea, the son of Young-Ja Shim and Sang-Woo Kim. After completing his graduation from Suncheon Hyocheon High School in 1988, he entered the Seoul National University. He received a B.A. in Education in 1995. He then spent six years working as a research assistant and a network engineer. In March 2002, he entered the graduate school of the Seoul National University to pursue a degree in educational counseling. He received his master's in educational counseling in February 2004. In August 2004, he entered the graduate school at the University of Texas at Austin to pursue a doctoral degree in Counseling Psychology in the Department of Educational Psychology. He will complete his pre-doctoral internship at the Florida State University Counseling Center in Tallahassee, Florida.

Permanent address: 2501 Lake Austin Blvd., Austin, TX 78703

This dissertation was typed by the author.